mccs

Datos disponibles en: www.sintesis.com

Análisis de datos en ciencias sociales y de la salud I

Antonio Pardo • Miguel Ángel Ruiz • Rafael San Martín

EDITORIAL SINTESIS

Análisis de datos
en ciencias sociales y de la salud I

PROYECTO EDITORIAL:

Metodología de las Ciencias del Comportamiento y de la Salud

Directores:

Antonio Pardo Merino
Miguel Ángel Ruiz Díaz

Análisis de datos

en ciencias sociales y de la salud I

Antonio Pardo • Miguel Ángel Ruiz • Rafael San Martín



Consulte nuestra página web: **www.sintesis.com**
En ella encontrará el catálogo completo y comentado

Vallehermoso, 34. 28015 Madrid
Teléfono 91 593 20 98
http://www.sintesis.com

ISBN: 978-84-995823-7-5

Impreso en España - Printed in Spain

Índice de contenidos

3. Análisis descriptivo de variables categóricas

4. Análisis descriptivo de variables cuantitativas

5. Puntuaciones típicas y curva normal

9. Inferencia con una variable

10. Inferencia con dos variables categóricas

11. Inferencia con una variable categórica y una cuantitativa

12. Inferencia con dos variables cuantitativas

Presentación

Este manual de análisis de datos es el primer volumen de una serie dedicada a revisar los procedimientos estadísticos comúnmente utilizados en el entorno de las ciencias sociales y de la salud.

Por qué un nuevo manual de "análisis de datos"

La decisión de los países del viejo continente de crear el Espacio Europeo de Educación Superior ha obligado a las universidades europeas a realizar una reforma generalizada de sus planes de estudios para adaptarlos a la nueva normativa. Esta reforma, que en muchos casos ha supuesto cambios importantes, ha afectado a todas las disciplinas; y los grados agrupados bajo la denominación de "ciencias sociales" y "ciencias de la salud" no son una excepción.

Este hecho, por sí sólo, ya bastaría para justificar la presentación de un nuevo manual de análisis de datos con contenidos adaptados a los nuevos grados. Pero lo cierto es que esto sólo ha sido la excusa para elaborar una propuesta acorde con nuestra forma de entender el análisis de datos o, quizá sería más exacto decir, acorde con nuestra idea acerca de cuál es la mejor manera de iniciar a un estudiante en el análisis de datos.

Qué contenidos seleccionar

Nuestra idea de cómo hacer las cosas afecta tanto a los contenidos seleccionados como a la forma de presentarlos. En la selección de los contenidos se ha tenido en cuenta, por un lado, la presencia cada vez más extendida de programas informáticos y de ordenadores donde poder utilizarlos; y, por otro, el hecho de que, por lo general, los estudiantes, profesores e investigadores que se mueven en el ámbito de las ciencias sociales y de la salud, ni son estadísticos ni pretenden serlo.

En primer lugar, el poder contar con programas informáticos capaces de aplicar cualquier procedimiento estadístico con suma facilidad y con el mínimo esfuerzo ha convertido en obsoletos algunos procedimientos a los que antes se les dedicaba bastate atención (por ejemplo, todo lo relativo a la agrupación de variables en intervalos o a los métodos abreviados de cálculo); dejar de lado estos procedimientos ha permitido liberar espacio para incluir otros nuevos. Pero además, ya no es necesario invertir tiempo haciendo a mano cálculos que no contribuyen en absoluto a entender el significado de lo que se está haciendo (como, por ejemplo, aplicar la fórmula clásica del coeficiente de correlación de Pearson para cuantificar el grado de relación lineal entre dos variables), lo cual contribuye de forma significativa a no tener que desviar la atención de lo realmente importante, que, en nuestra opinión, no es precisamente

realizar cálculos, sino aprender a elegir el procedimiento apropiado en cada caso y a interpretar correctamente los resultados que ofrece. Aunque todos los procedimientos se presentan con suficiente detalle como para poder aplicarlos a mano, de todos ellos se explica también cómo aplicarlos con un programa informático.

En segundo lugar, no se presta atención detallada a algunos contenidos que, por ser fundamento matemático de los procedimientos estadísticos, es habitual encontrarlos en la mayoría de los manuales sobre análisis de datos y estadística aplicada. En este sentido, no se ofrece ningún capítulo dedicado a contenidos que suelen recibir bastante atención en otros manuales: la teoría de la probabilidad, que suele merecer al menos un capítulo, se resume en un apartado; las distribuciones de probabilidad, que suelen tratarse como un bloque en uno o dos capítulos, como un listado de distribuciones independientes de lo demás, aquí se presentan como parte complementaria de otros procedimientos y sólo se presta atención a las más importantes; y todo lo relacionado con el estudio pormenorizado de la variables aleatorias (valores esperados y momentos, funciones de probabilidad, etc.) se ha reducido a la mínima expresión. Un profesional de las ciencias sociales y de la salud no es un estadístico; y, muy probablemente, tampoco pretende serlo; consecuentemente, no necesita ser un experto en los fundamentos matemáticos de las herramientas estadísticas. Creemos que el énfasis hay que colocarlo, más bien, en conocer la utilidad de los procedimientos disponibles y en saber elegirlos, aplicarlos e interpretarlos correctamente.

Cómo presentar los contenidos

En nuestra idea acerca de cuál es la mejor manera de iniciar a un estudiante en el análisis de datos también desempeña un papel importante la forma de presentar los contenidos. Y esto afecta tanto a la ordenación de los mismos como a la forma de exponerlos.

En lo relativo a la ordenación de los contenidos, nuestra opción, aunque poco convencional, nos ha parecido que era la mejor apuesta. Por lo general, los manuales de introducción al análisis de datos presentan una primera parte con herramientas descriptivas para una variable (distribuciones de frecuencias, gráficos, estadísticos de posición, dispersión y forma, etc.) y para dos o más variables (distribuciones conjuntas, relación lineal, regresión lineal), dedicando una segunda parte a la lógica de la inferencia estadística y a algunas herramientas inferenciales concretas para una y dos variables (con atención ocasional al estudio de más de dos variables). En nuestra propuesta, las herramientas descriptivas se explican referidas a una sola variable; después se introducen los conceptos inferenciales y, a partir de ahí, el análisis de dos o más variables se realiza mezclando las herramientas descriptivas con las inferenciales.

Varias razones nos han hecho optar por este formato. La primera es más bien de tipo, podríamos decir, *profesional*. Cuando un analista de datos se enfrenta con un archivo de datos, la primera tarea que aborda es la de intentar formarse una idea lo más exacta posible acerca de las características de los datos. Esta tarea se lleva a cabo aplicando, sin ideas preconcebidas, herramientas descriptivas y exploratorias a cada una de las variables individualmente consideradas intentando identificar tanto regularidades como anomalías. En una segunda fase, el analista mezcla variables para obtener nueva información sobre las características de los datos. Pero cuando se estudian dos o más variables simultáneamente es porque interesa, no sólo describirlas, sino compararlas o relacionarlas de acuerdo con un plan preconcebido (no se mezcla todo con todo); y para esto no suele ser suficiente aplicar herramientas descriptivas;

son las herramientas inferenciales las que ayudan a detectar la posible presencia de diferencias o relaciones. De ahí que, en el estudio de más de una variable, nos parezca conveniente estudiar simultáneamente las herramientas descriptivas y las inferenciales.

La segunda razón es de tipo *docente*. Según nuestra experiencia, los conceptos inferenciales y la lógica en la que se basan son los contenidos que más cuesta asimilar a quienes se acercan por primera vez al análisis de datos. Para facilitar la comprensión de estos contenidos nos ha parecido buena idea darles dos repasos: uno más básico e intuitivo (en este primer volumen) y otro algo más profundo y fundamentado (en el segundo volumen). De ahí que en este primer volumen hayamos decidido incluir varios capítulos dedicados a la lógica de la inferencia estadística y a la aplicación de algunas herramientas inferenciales concretas.

Por lo que se refiere a la forma de exponer los contenidos, se ha intentado prestar más atención a los aspectos prácticos o aplicados que a los teóricos o formales, pero sin descuidar estos últimos. Aunque este manual va dirigido, principalmente, a estudiantes de disciplinas del ámbito de las ciencias sociales y de la salud, no se trata de un material diseñado exclusivamente para ellos. También pretende servir de ayuda a los profesores de análisis de datos y a los investigadores. Creemos que ambos pueden encontrar, en éste y en los siguientes volúmenes, las respuestas a muchas de las preguntas que se hacen en su trabajo cotidiano.

Y todo ello sin olvidar que, en los tiempos que corren, no tiene sentido analizar datos sin el apoyo de un programa informático. Ahora bien, conviene tener muy presente que, aunque las herramientas informáticas pueden realizar cálculos con suma facilidad, todavía no están capacitadas para tomar algunas decisiones. Un programa informático no sabe si la estrategia de recogida de datos utilizada es la correcta, o si las mediciones aplicadas son apropiadas; tampoco decide qué prueba estadística conviene aplicar en cada caso, ni interpreta los resultados del análisis. Los programas informáticos todavía no permiten prescindir del analista de datos. Es él, el analista, quien debe mantener el control de todo el proceso. El éxito de un análisis depende de él y no del programa informático. El hecho de que sea posible ejecutar las técnicas de análisis más complejas con la simple acción de pulsar un botón sólo significa que es necesario haber atado bien todos los cabos del proceso (diseño, medida, análisis, etc.) antes de pulsar ese botón.

No podemos dejar pasar la oportunidad que nos brinda esta presentación para agradecer a nuestro compañero Ludgerio Espinosa, y a muchos de nuestros alumnos y a no pocos lectores de nuestros trabajos previos, las permanentes sugerencias hechas para mejorar nuestras explicaciones y la ayuda prestada en la caza de erratas. Los errores y deficiencias que todavía permanezcan son, sin embargo, atribuibles sólo a nosotros.

Antonio Pardo
Miguel Ángel Ruiz
Rafael San Martín

1

Introducción al análisis de datos

Qué es el análisis de datos

El *análisis de datos*, o *análisis estadístico*, es un conjunto de procedimientos diseñados para *resumir y organizar datos con el objetivo de extraer información y elaborar conclusiones*. Este conjunto de procedimientos (del que todas las ciencias empíricas –medicina, biología, psicología, sociología, economía, educación, etc.– hacen uso) pertenece a una rama de las matemáticas conocida con el nombre de *estadística*. Esta moderna ciencia, la estadística, es el resultado de la confluencia de dos disciplinas independientes: el *cálculo de probabilidades*, que surge como una aproximación matemática a los juegos de azar, y la *estadística*, o ciencia del Estado, dedicada a llevar registros ordenados (contar, tabular, clasificar, censar, etc.) de los datos del Estado. La unión de ambas en el siglo XIX dio lugar a una nueva ciencia interesada, fundamentalmente, en estudiar cómo obtener conclusiones de la investigación empírica mediante el uso de modelos matemáticos (ver Hays, 1994). La estadística puede concebirse como una *ciencia que recoge, ordena y analiza los datos de una muestra extraída de una determinada población, para hacer inferencias acerca de esa población valiéndose del cálculo de probabilidades* (Amón, 1979, pág. 37).

Es común encontrar la estadística dividida en dos partes: *descriptiva* e *inferencial*. La *estadística descriptiva* consta de una serie de procedimientos diseñados para describir la información contenida en un conjunto de datos (generalmente una muestra); es lo que se corresponde con lo que hemos llamado *resumir y organizar datos*.

La *estadística inferencial* (también llamada *inductiva*) engloba una serie de procedimientos que permiten generalizar (inferir, inducir) la información contenida en ese conjunto particular de datos (muestra) al conjunto total de datos (población) a los que representan; es lo que se corresponde con lo que hemos llamado *extraer información y elaborar conclusiones*.

Por supuesto, para poder efectuar este salto de lo particular a lo general es crucial que el conjunto de datos utilizados para obtener información (muestra) sea *representativo* del conjunto total de datos sobre el que se desea realizar la inferencia (población); es decir, es necesario efectuar una correcta *selección* de los datos. Esto se consigue mediante técnicas de *muestreo*, las cuales también pertenecen al ámbito de la estadística.

En ocasiones se habla del *cálculo de probabilidades* como de una parte de la estadística, pero quizá es preferible entenderlo como una herramienta matemática de la que se sirve la

estadística. En estadística descriptiva, el cálculo de probabilidades ayuda a mejorar nuestro conocimiento acerca de cómo se comportan los datos; en estadística inferencial se utiliza para realizar el salto (hacer inferencia) de lo particular a lo general.

Una distinción habitual especialmente apropiada para contextualizar los contenidos de este manual es la que se da entre estadística *teórica* y estadística *aplicada*. La estadística teórica se ocupa de investigar y proponer métodos formalmente válidos para organizar los datos y realizar inferencias; la estadística aplicada se ocupa de aplicar esos métodos a datos reales. Esta versión aplicada es lo que actualmente se conoce (probablemente gracias a Tukey, 1962) como **análisis de datos** (Agresti, 2002, 2007; Hardy y Bryman, 2004; Keppel y Wickens, 2004; Keren y Lewis, 1993; Linoff, 2008; Maxwell y Delaney, 2004; Ramsay y Silverman, 2002; Tamhane y Dunlop, 2000; Velleman y Hoaglin, 2004; Zuur, Ieno y Smith, 2007; etc.).

En general, el análisis de datos es un proceso que se desarrolla en fases: empieza con la selección y recopilación de los datos (tarea que debe estar guiada por los objetivos del estudio y por el correspondiente diseño de investigación), continúa con la aplicación de herramientas descriptivas para organizar y resumir la información contenida en los datos y termina (no necesariamente, pero sí frecuentemente) con la aplicación de herramientas inferenciales para efectuar comparaciones y estudiar relaciones.

Para qué sirve el análisis de datos

Las ciencias pueden clasificarse en *formales* y *empíricas*. En las ciencias formales (las matemáticas, por ejemplo) no hay necesidad de entrar en contacto con el mundo real; basta con establecer un conjunto de postulados y proceder a partir de ellos por deducción lógica.

En las ciencias empíricas, por el contrario, el objetivo fundamental es el de encontrar relaciones de tipo general (leyes) capaces de explicar el comportamiento de uno o varios eventos reales cuando se dan las circunstancias apropiadas. Y, a diferencia de lo que ocurre en las ciencias formales, esas leyes sólo pueden ser descubiertas y verificadas observando el mundo real. Sin embargo, no existe científico o grupo de científicos capaces de observar todos los posibles eventos relacionados con una determinada ley. Las conclusiones sobre lo que ocurrirá con la totalidad de una clase particular de eventos se extraen a partir de la observación de sólo unos pocos eventos de esa clase. Esto es lo que se conoce como *inducción* o generalización inductiva.

Mientras las leyes de la deducción lógica (propias de las ciencias formales) permiten llegar a conclusiones verdaderas a partir de premisas verdaderas, la generalización inductiva (propia de las ciencias empíricas) intenta ir desde lo que se observa que ocurre en un *conjunto* reducido de observaciones hasta la afirmación de que eso mismo es válido también para el *total* de observaciones de la misma clase.

Este *salto* de lo particular a lo general posee un *riesgo* nada despreciable. Multitud de factores influyen sobre los eventos observables alterando las similitudes y diferencias entre ellos. Podría decirse que cada observación es, en algún sentido, diferente de la siguiente. En ciencias como la física (en algunas de sus parcelas, al menos), esta diferencia entre observaciones consecutivas es, generalmente, bastante reducida, de modo que unas pocas observaciones de un mismo evento suelen producir resultados muy parecidos, si no idénticos. En este contexto, la generalidad de las conclusiones obtenidas inductivamente no constituye un problema importante. Pero ése no es el caso en la mayoría de las ciencias empíricas (medicina, biología, psicología, sociología, economía, etc.). En estas ciencias, la variación existente entre las dis-

tintas observaciones de un mismo evento no puede ser sometida, habitualmente, a un control riguroso. Las fuentes de variación existentes son muy numerosas y resultan extremadamente difíciles de identificar, medir y controlar. En estas circunstancias, las conclusiones a las que es posible llegar inductivamente requieren la utilización de una metodología en cierto sentido especial. Y es precisamente la estadística, mediante el conjunto de procedimientos o herramientas englobadas bajo la denominación de *análisis estadístico* o *análisis de datos*, la encargada de proporcionar a las ciencias empíricas esa metodología.

La más importante aplicación del análisis de datos está, por tanto, relacionada con el concepto de *incertidumbre*, entendida ésta como la tendencia de un resultado a variar cuando se efectúan repetidas observaciones del mismo bajo condiciones idénticas. En situaciones *deterministas*, en las que una misma causa produce siempre un mismo resultado (un cuerpo desplazado a una velocidad constante v durante un tiempo t recorre un espacio e), el álgebra o el análisis matemático bastan para alcanzar el nivel de comprensión buscado. Por el contrario, en situaciones *aleatorias*, en las que una misma causa puede producir cualquiera de un conjunto de resultados posibles (lanzar una moneda al aire, observar la respuesta de un paciente a un tratamiento, etc.), es necesario recurrir al análisis de datos (a las herramientas que proporciona la estadística) para poder extraer conclusiones fiables.

Niveles de indagación: descriptivo, relacional, explicativo

Ya hemos señalado que el análisis de datos debe ser entendido, ante todo, como un conjunto de herramientas al servicio de la investigación empírica. Ahora bien, la investigación empírica puede desarrollarse en diferentes niveles de *indagación* (puede consultarse Rosenthal y Rosnow, 1991, para profundizar en los contenidos de este apartado).

Supongamos que un investigador interesado en comprender ciertos aspectos relacionados con el rendimiento académico viene observando que entre los alumnos de enseñanza primaria existen diferencias individuales en comprensión lectora. Para obtener alguna evidencia adicional sobre su sospecha, decide seleccionar una muestra aleatoria de sujetos y pasarles una prueba de comprensión lectora. Supongamos que, analizados los datos, nuestro investigador encuentra que los sujetos, efectivamente, difieren en comprensión lectora. Su investigación se encuentra, de momento, en un nivel al que podemos llamar **descriptivo**: ha conseguido dar respuesta a la pregunta *cómo son las cosas* (en concreto, ha encontrado que no todos los sujetos tienen el mismo nivel de comprensión lectora). En este nivel de indagación se intenta obtener conocimiento sobre algo desconocido, identificar problemas de investigación y generar ideas (posibles soluciones a los problemas) para ser estudiadas a otros niveles.

Constatado el hecho de que los sujetos difieren en comprensión lectora, nuestro investigador decide averiguar si esos mismos sujetos difieren en el tipo de pautas motivacionales que manifiestan. Evalúa tal circunstancia (existen procedimientos apropiados para ello; ver Pardo y Alonso, 1990) y llega a la conclusión de que, efectivamente, los sujetos muestran pautas motivacionales diferentes.

Si nuestro investigador decidiera detener ahí su estudio, se quedaría ubicado en un nivel de indagación de tipo descriptivo. Pero este nivel de indagación raramente resulta satisfactorio para un investigador curioso. Por esta razón, decide poner en relación los dos hechos observados y descubre que en los sujetos con mejor comprensión lectora predomina un tipo de pautas motivacionales (orientación hacia el aprendizaje) completamente diferentes de las que predominan en los sujetos con peor comprensión lectora (orientación hacia la ejecución).

Nuestro investigador se ha situado en un segundo nivel de indagación que suele denominarse **relacional**: ha conseguido dar respuesta a la pregunta *cómo unas cosas se relacionan con otras*. Es razonable pensar que el resultado de la investigación empírica no puede limitarse únicamente a una colección de hechos. Los hechos deben ser conectados entre sí de una forma lógica y sistemática para constituir conocimiento organizado. La investigación de tipo relacional permite avanzar hacia ese objetivo intentando (1) descubrir qué hechos se encuentran relacionados –y en qué medida– y (2) predecir unos a partir de otros.

Supongamos por último que nuestro investigador, sospechando que las pautas motivacionales específicas de cada sujeto podrían estar mediatizando el nivel de comprensión lectora que alcanzan, decide seleccionar dos grupos aleatorios de sujetos y entrenar a cada uno con pautas motivacionales diferentes (existen procedimientos apropiados para ello; ver Pardo y Alonso, 1990). Finalizado el entrenamiento, evalúa el nivel de comprensión lectora de todos los sujetos y encuentra que los entrenados en orientación hacia el aprendizaje muestran mejor comprensión lectora que los entrenados en orientación hacia la ejecución.

Nuestro investigador se acaba de situar en un nivel de indagación denominado **explicativo**: ha conseguido dar respuesta (al menos una respuesta parcial) a la pregunta *cómo las cosas han llegado a ser de la forma que son* o, dicho de otra forma, *por qué las cosas son como son*. Este nivel de indagación permite establecer relaciones de tipo causal entre los eventos, de manera que lo que ocurre con uno o varios de ellos puede ser explicado recurriendo a otro o varios diferentes: por ejemplo, las pautas motivacionales *influyen* sobre el nivel de comprensión lectora; es decir, los sujetos difieren en comprensión lectora *porque* poseen, entre probablemente otras cosas, pautas motivacionales diferentes.

Esta distinción entre niveles de indagación es de fundamental importancia a la hora de establecer el tipo de conclusiones que es posible extraer de un análisis de datos. Por supuesto, las técnicas de análisis de datos pueden ser aplicadas a cualquier conjunto de datos independientemente del nivel de indagación en el que pueda situarse un estudio (es decir, en todo estudio se dispone de datos susceptibles de ser analizados). Pero una técnica de análisis de datos no determina el nivel de indagación en el que es posible situar las conclusiones de un estudio. Esto viene condicionado, no por la técnica de análisis, sino por la estrategia de *recogida de datos* adoptada.

La recogida de datos se realiza en el contexto de un **diseño de investigación** (es decir, de un plan de actuación) y, dependiendo del fenómeno estudiado y del nivel de comprensión que se desee (o se pueda) obtener del mismo, puede efectuarse siguiendo dos caminos alternativos: (1) esperando que aparezca el fenómeno que se desea estudiar y observándolo cuando ocurre (diseño **observacional**); y (2) provocando que ocurra bajo circunstancias controladas y registrándolo al producirse (diseño **experimental**). Estas dos formas alternativas de plantear la recogida de datos (con las variantes que se quieran añadir) difieren, básicamente, en el grado de *control* que se ejerce sobre los diferentes elementos de la situación en la que se da el fenómeno que se desea estudiar, siendo este control máximo en la metodología experimental y mínimo en la observacional.

En la metodología experimental se ejerce control tanto sobre las condiciones del estudio como sobre el conjunto de variables no incluidas en el estudio; las condiciones del estudio se controlan creándolas (es el propio investigador el que genera las condiciones; por ejemplo, el tipo de tratamiento, la intensidad de un estímulo; etc.); el resto de variables se controlan asignando a cada condición del estudio una muestra aleatoria de sujetos: se asume que el azar iguala los grupos en el conjunto de variables no incluidas en el estudio. Cuando el investigador de nuestro ejemplo se encontraba en el nivel descriptivo, se había limitado a seleccionar

una muestra aleatoria de sujetos y a obtener un registro de la respuesta que deseaba estudiar sin ejercer control sobre ningún elemento de la situación. Posteriormente, al situarse en el nivel explicativo, ejerció control sobre el tipo de pautas motivacionales (manipuló las condiciones del estudio creando dos tipos de entrenamiento: orientación al aprendizaje y orientación a la ejecución) e igualó los grupos asignando los sujetos aleatoriamente. Además, creó una situación en la que se podrían haber controlado otros aspectos (igualando los dos grupos en el nivel de comprensión lectora previa al entrenamiento o en alguna otra característica sospechosa de afectar a la comprensión lectora como, por ejemplo, el cociente intelectual, etc.).

A medio camino entre la metodología observacional y la experimental se encuentra la metodología **correlacional** (también llamada *selectiva* y *cuasi-experimental*). Se da en ella un mayor grado de control que en la observacional (existe, por ejemplo, selección –de ahí el nombre de selectiva– de las condiciones del estudio y, en ocasiones, manipulación de esas condiciones –de ahí el nombre de cuasi-experimental–; es posible controlar el efecto de algunas variables no incluidas en el estudio; etc.), pero no se da en ella el grado de control propio de la metodología experimental. La principal diferencia entre ambas metodologías es que en la correlacional no existe asignación aleatoria de los sujetos a las condiciones del estudio. Esto se debe, unas veces, a que no es posible hacerlo; por ejemplo, al comparar el salario de hombres y mujeres no es posible decidir quién es hombre y quién es mujer; eso es algo que viene dado. Y, otras, a que no interesa hacerlo por razones prácticas o éticas; por ejemplo, al comparar dos métodos de enseñanza se decide aplicar cada método a los alumnos de un aula simplemente porque no se considera apropiado mezclar los alumnos aleatoriamente sólo por el interés de la investigación; o al comparar los problemas respiratorios de fumadores y no fumadores no se considera ético (ni probablemente habría sujetos dispuestos a ello) seleccionar al azar sujetos no fumadores y convertir a parte de ellos en fumadores.

La principal consecuencia de la falta de asignación aleatoria a las condiciones del estudio es que no se trabaja con grupos equivalentes. Y esto significa que, a diferencia de lo que ocurre con los diseños experimentales, con los selectivos y cuasi-experimentales no es posible conocer con certeza si las diferencias y relaciones encontradas se deben a las condiciones del estudio (sexo, tabaquismo, etc.) o a la presencia de variables que no se han podido controlar.

Lo que interesa destacar al introducir esta breve descripción de las diferentes estrategias de recogida de datos es que la utilización de una u otra técnica de análisis (es decir, de una u otra herramienta estadística) no determina, por ella misma, el tipo de conclusiones que es posible extraer de un análisis. Una técnica de análisis sirve para comparar grupos y para relacionar variables y, como consecuencia de ello, para detectar posibles diferencias y posibles relaciones. Pero una técnica de análisis no permite saber a qué se deben las diferencias y relaciones detectadas. Ciertamente, hay algunas técnicas de análisis más características de unas metodologías que de otras. Pero el principal determinante del nivel de indagación en el que es posible situarse no es la técnica de análisis aplicada, sino la estrategia de recogida de datos utilizada. En términos generales, se puede afirmar que las metodologías observacional y selectiva generan, básicamente, investigación descriptiva y relacional, mientras que la metodología experimental genera, básicamente, investigación explicativa[1].

---

[1] La posibilidad de establecer relaciones de tipo *causal* entre eventos no es algo que venga determinado exclusivamente (aunque tal vez sí principalmente) por la estrategia de recogida de datos utilizada. Cuando un cuerpo de conocimientos bien organizado (teoría) es capaz de predecir determinado tipo de estructura relacional entre eventos, también es posible llegar a conclusiones de tipo causal (independientemente del nivel de indagación alcanzado debido a las restricciones impuestas por el diseño de investigación). Para profundizar en esta problemática puede consultarse, por ejemplo, Davis (1985).

Escalas de medida

El análisis de datos se basa, obviamente, en *datos*. Pero un dato no es otra cosa que un *número*. Lo cual significa que, para poder analizar datos, es necesario asignar números a las características de las personas u objetos que se desea estudiar. Ahora bien, ese proceso consistente en asignar números a las características que se desea estudiar, proceso denominado *medida* o *medición*, no es responsabilidad de la estadística. De ese proceso se encarga la *teoría de la medida*, la cual tiene por objeto el estudio de los diferentes modelos que permiten establecer las reglas que es necesario seguir para una correcta asignación de números.

Si la característica o propiedad que se desea medir existe en una cierta cantidad, la medición consiste simplemente en asignar a esa característica, de acuerdo con alguna regla, un número capaz de expresar esa cantidad con la mayor precisión posible. Así es como se hace con características como la longitud, el peso o el tiempo; disponiendo de un instrumento de medida apropiado, la medición de estas características no constituye un problema importante. El problema surge cuando se desea medir características que no está tan claro que puedan ser cuantificadas.

No es éste, por supuesto, el lugar adecuado para entrar en el debate histórico que ha suscitado este problema, pero sí nos parece conveniente señalar que, gracias al persistente esfuerzo de muchos científicos, a partir del Congreso sobre Medición para el Avance de la Ciencia y la Tecnología, celebrado en Moscú en 1979, la medición de características aparentemente poco cuantificables dejó de ser prohibitiva y empezó a adquirir el reconocimiento por el que tanto tiempo estuvo luchando.

Hoy, la medición no se concibe exactamente como la asignación de un numeral que exprese la magnitud de cierta propiedad. Medir consiste en hacer corresponder dos sistemas de relaciones: uno *empírico* (el de las propiedades que se desea medir) y otro *formal* (el de los números que se asignan en la medición). Es necesario que las relaciones presentes en el sistema formal reflejen las presentes en el sistema empírico para que la correspondencia efectuada se considere una medición.

Consideremos, como ejemplo, la característica *sexo*. Para analizar datos referidos a esa característica puede atribuirse el número 1 a la modalidad *hombre* y el número 2 a la modalidad *mujer*. Consideremos ahora dos individuos y la característica sexo. O los dos individuos son hombres, o los dos son mujeres, o uno es hombre y el otro mujer. Desde el punto de vista del análisis de datos, tras la medición se tendrán dos unos, dos doses, o un uno y un dos. Ahora bien, entre esos números sólo podrá establecerse una relación de igualdad o desigualdad. No se podrá, por ejemplo, establecer una relación de orden (es decir, de *mayor* o *menor*), pues el valor 2 no indica mayor *cantidad* que el valor 1 (ser mujer no indica, como es obvio, mayor posesión de la característica sexo que ser hombre, a pesar de que $1 < 2$).

En este caso, los números sólo sirven para identificar o distinguir las dos modalidades de la característica sexo. Sin embargo, en otros casos, con otras características, los números permiten establecer otro tipo de relaciones. Los números que se asignan a la característica *altura*, por ejemplo, reflejan relaciones diferentes de las que reflejan los asignados a la característica sexo. Un individuo que mide 180 cm posee más *cantidad* de altura que otro sujeto que mide 160 cm. Es decir, no todas las características se miden de la misma forma (los números que se asignan no siempre significan lo mismo) porque entre sus valores no siempre se da el mismo tipo de relación. La medición es en unos casos *mejor* que en otros, en el sentido de que en unos casos permite establecer mayor número de relaciones que en otros.

Clasificación de Stevens

De lo anterior cabe deducir que, dependiendo del tipo de relaciones que puedan establecerse entre los valores (números) asignados a una característica, es posible definir diferentes *niveles* o *escalas de medida*. Tradicionalmente se han distinguido cuatro niveles o escalas de medida (Stevens, 1946, 1951): nominal, ordinal, de intervalos y de razón (utilizaremos indistintamente los términos *escalas* y *niveles* de medida; por tanto, de una característica medida utilizando, por ejemplo, una escala ordinal, podremos decir, queriendo significar lo mismo, que su nivel de medida es ordinal).

La medida **nominal** consiste en clasificar en categorías a los sujetos u objetos que se desea medir haciendo que todos los sujetos u objetos clasificados dentro de la misma categoría sean equivalentes en la característica que se está midiendo. Tras esto, se asignan números a las categorías establecidas y se considera que todos los sujetos u objetos a los que se les ha asignado el mismo número son cualitativamente *iguales* en la variable medida, mientras que los sujetos u objetos a los que se les ha asignado un número diferente (por haber sido clasificados en categorías diferentes) se considera que son cualitativamente *distintos*. Las categorías utilizadas (tantas como valores o niveles distintos tenga la característica que se desea medir) deben reunir dos propiedades: *exhaustividad* (todos los sujetos u objetos pueden ser clasificados en alguna de las categorías establecidas) y *exclusividad* (cada sujeto u objeto puede ser clasificado en sólo una de las categorías establecidas; es decir, las categorías no se solapan).

Esta escala de medida es la más *débil* de todas en el sentido de que la única relación que es posible establecer entre los sujetos u objetos medidos es la de *igualdad-desigualdad*. Los números asignados actúan simplemente como nombres o etiquetas para identificar las categorías establecidas: en lugar de números podría utilizarse cualquier otro símbolo y nada cambiaría (quizá no debería decirse que la escala nominal es una escala, pues no se está escalando nada; simplemente se están asignando etiquetas a los sujetos u objetos medidos).

Existen muchas características en las que sólo puede conseguirse un nivel de medida nominal: el sexo (masculino, femenino), el estado civil (soltero, casado, divorciado, etc.), el lugar de procedencia (Madrid, Galicia, Andalucía, etc.), la nacionalidad, la raza, el tipo de enfermedad, el tipo de tratamiento, el resultado de una tarea (éxito, fracaso), la actitud hacia un objeto (a favor, en contra), etc. Para poder utilizar el análisis de datos con estas variables es necesario asignar un valor numérico a cada uno de sus valores. Si deseamos medir, por ejemplo, la variable *tipo de neurosis*, podemos asignar un 1 a los sujetos con *neurosis obsesiva*, un 2 a los sujetos con *neurosis histérica*, un 3 a los sujetos con *neurosis fóbica*, etc. Pero es obvio que, viendo de qué tipo de variable se trata, los números asignados serán, a todos los efectos, meros rótulos, por lo que lo único que permitirán afirmar acerca de los sujetos u objetos medidos es si son iguales o distintos en la variable medida, es decir, si pertenecen o no a la misma categoría de la variable (obviamente, un sujeto con neurosis fóbica *no es igual* a un sujeto con neurosis obsesiva *más* otro con neurosis fóbica, a pesar de que $3 = 1 + 2$; y esta igualdad no se verifica porque la asignación de los valores 1, 2, 3, ..., se ha hecho de forma arbitraria).

La medida **ordinal** consiste en asignar a los sujetos u objetos medidos un número que permita *ordenarlos* según la cantidad que poseen de la característica medida. En la escala o medida ordinal, además de estar presente la relación de igualdad-desigualdad propia de la escala nominal, los números asignados permiten saber si la cantidad de característica que posee un sujeto u objeto es *mayor que* o *menor que* la cantidad que posee otro sujeto u objeto cualquiera.

En las ciencias sociales y de la salud es frecuente encontrarse con características en las que resulta apropiado utilizar una escala de medida ordinal. La satisfacción con un producto o servicio, el bienestar psicológico, el dolor percibido, la actitud hacia sujetos u objetos, el cociente intelectual, etc., son ejemplos de características que suelen medirse con una escala ordinal. Es posible ordenar, por ejemplo, un conjunto de sujetos según su grado de satisfacción con un determinado servicio: asignando un 1 al sujeto más satisfecho, un 2 al sujeto más satisfecho de los restantes, un 3 al siguiente, etc. Al final se tendrán *n* sujetos ordenados según su grado de satisfacción. Al hacer esto, ya no sólo es posible afirmar que dos sujetos a los que se les ha asignado un número diferente poseen un grado de satisfacción diferente (como se hace en el nivel de medida nominal), sino, además, que el grado de satisfacción de tal sujeto es mayor o menor que el de tal otro. Sin embargo, no es posible afirmar nada acerca de la magnitud de la diferencia existente entre dos números. En el nivel de medida ordinal se desconoce si la diferencia existente entre los sujetos a los que se les ha asignado un 1 y un 2 es igual (o distinta) que la diferencia existente entre los sujetos a los que se les ha asignado un 3 y un 4. De modo que la diferencia en grado de satisfacción entre los sujetos a los que se les ha asignado un 1 y un 2 puede no ser (y normalmente, en este nivel de medida, no lo será) la misma que entre los sujetos a los que se les ha asignado un 2 y un 3.

En la medida de **intervalos**, además de poder afirmar que la característica medida se da con mayor intensidad en un sujeto u objeto que en otro (relación alcanzada ya en la escala ordinal), también es posible determinar la magnitud de la diferencia existente entre dos sujetos u objetos medidos, es decir, la cantidad en la que difieren.

Se elige una unidad de medida y, tras ello, se asigna a cada sujeto u objeto medido un número indicativo de la cantidad de característica que posee en términos de las unidades de medida elegidas. Así, un objeto al que se le asigna la puntuación 12 en una escala de intervalos tiene 2 unidades de medida más que un objeto al que se le asigna la puntuación 10; del mismo modo, un objeto al que se le asigna la puntuación 6 tiene 2 unidades de medida más que un objeto al que se le asigna la puntuación 4. Entre 10 y 12 existe la misma diferencia, en cantidad de característica, que entre 4 y 6.

Pero en la escala de intervalos no se puede afirmar que 12 es el doble de 6. La razón es que no existe *cero absoluto*, es decir, no existe un valor numérico que indique ausencia total de cantidad. El valor numérico 0 es un punto más de la escala, un punto arbitrario. La temperatura, por ejemplo, es una característica que se mide utilizando una escala de intervalos. Cuando se dice, en escala Celsius, que ayer hubo 20 grados de temperatura máxima y hoy 25, se está diciendo no sólo que hoy hubo más temperatura que ayer (afirmación propia de la escala ordinal), sino que hoy hubo 5 grados más de temperatura que ayer. Del mismo modo, 20 grados son 5 más que 15. La diferencia entre 15 y 20 grados es la misma que entre 20 y 25, y esto va más allá de lo que puede afirmarse con una escala ordinal. Sin embargo, no es posible afirmar que 20 grados representen el doble de temperatura que 10, pues en la escala Celsius, el valor cero es un punto arbitrario de la escala y, por tanto, no indica ausencia de la característica medida.

La medida de **razón** añade a la de intervalos la presencia del cero absoluto. Es decir, el cero de una escala de razón indica ausencia total de la característica medida. Por tanto, el cero no es un punto arbitrario de la escala (como ocurre en la escala de intervalos), sino el punto que indica que no existe cantidad alguna de la característica medida. Al igual que en la escala de intervalos, también aquí las diferencias entre los objetos medidos son constantes (existe una unidad de medida), pero, además, la presencia del cero absoluto permite afirmar que la característica medida se da en un objeto el doble, el triple, etc., que en otro.

El tiempo, la extensión y el peso, por ejemplo, son características medidas en escala de razón. No sólo es posible afirmar que la diferencia existente entre 30 y 60 segundos es la misma que entre 60 y 90 (afirmación válida también en la escala de intervalos), sino, además, que 60 segundos son el doble de 30 segundos. Además, el valor cero referido al tiempo, o a la extensión, o al peso, indica que no hay tiempo, o extensión, o peso.

Rol de las escalas de medida

La importancia de distinguir apropiadamente las diferentes escalas de medida radica en que la elección de los diferentes procedimientos estadísticos está, en buena medida, condicionada por el tipo de mediciones de que se dispone. Éste es el punto de vista defendido por muchos autores a partir de las aportaciones iniciales de Stevens (Kornbrot, 1990; Siegel y Castellan, 1988; Townsend y Ashby, 1984; Westerman, 1983; etc.). Para otros muchos, sin embargo, el rol de las escalas de medida en la elección de los procedimientos estadísticos es más bien irrelevante (Anderson, 1961; Binder, 1984; Davison y Sharma, 1988; Gaito, 1960, 1980, 1986; Gardner, 1975; Maxwell y Delaney, 1985; Wright, 1997a, etc.). Para una revisión de los distintos puntos de vista relacionados con esta controversia puede consultarse Zumbo y Zimmerman (2000). Aunque éste no es el lugar para profundizar en esta problemática, sí nos parece conveniente hacer algunas reflexiones que pueden resultar útiles para el analista de datos.

Existen multitud de características de diferente índole en las que no resulta fácil determinar el nivel de medida que es posible alcanzar. El hecho de que las cuatro escalas de medida descritas sean exhaustivas (cualquier característica puede ser medida con alguna de ellas) y mutuamente exclusivas[2] (no se solapan), constituye un verdadero problema a la hora de trabajar con algunas características.

Supongamos que se mide la característica *percepción subjetiva de dolor* en 3 sujetos con una escala de 0 a 100 puntos, y que se obtiene una puntuación de 10 para el primero de ellos, de 20 para el segundo y de 90 para el tercero. En sentido estricto, en una escala de este tipo no es posible afirmar que la distancia existente entre una puntuación de 10 y otra de 20 (10 puntos) es equivalente a la distancia existente entre una puntuación de 50 y otra de 60 (también 10 puntos). Y no es posible afirmar tal cosa porque en una escala de percepción subjetiva de dolor no existe una unidad de medida que garantice tal equivalencia. Según esto, debería considerarse que la medida obtenida es de tipo ordinal, lo que permitiría concluir, tan sólo, que el tercer sujeto percibe *más dolor* que el segundo y éste más que el primero. Sin embargo, si se pidiera opinión a unos cuantos expertos, seguramente todos coincidirían en afirmar, no sólo que el tercer sujeto (90) manifiesta percibir más *dolor* que los otros dos (10 y 20), sino que las respuestas de estos dos sujetos se parecen entre sí más de lo se parecen a la respuesta del tercero. Ahora bien, esta afirmación supera el alcance de las propiedades de una escala ordinal. Consecuentemente, parece razonable pensar que una escala de percepción subjetiva de dolor (al igual que otras muchas: de actitudes, de satisfacción, de calidad de vida, de depre-

[2] Unas escalas de medida son más *débiles* o más *fuertes* que otras en función de la precisión con la que permiten medir (la escala nominal es la más débil; la escala de razón es la más fuerte). Y las propiedades de una escala de medida inferior (más débil) están contenidas en cualquiera de las escalas superiores (más fuertes). Por tanto, parece claro que las diferentes escalas de medida no pueden considerarse, en rigor, mutuamente exclusivas. Sin embargo, si atendemos al nivel de medida más alto que puede alcanzar una característica en función del tipo de relaciones que puedan establecerse entre sus niveles, entonces sí es posible hablar de exclusividad, pues, desde este punto de vista, a una característica dada sólo le corresponde un nivel de medida.

sión, de inteligencia, etc.) no puede identificarse con la escala ordinal común. Más bien parece que la frontera entre la escala ordinal y la de intervalos es lo bastante difusa como para poder asegurar que muchas características aparentemente medidas con una escala ordinal pueden ser tratadas como si estuvieran medidas con una de intervalos. Y esto es algo que un analista de datos no puede pasar por alto.

Para terminar este apartado conviene insistir en una idea importante. En principio, cualquier conjunto de números es susceptible de ser tratado por cualquiera de las herramientas estadísticas disponibles; es decir, no existe ninguna herramienta estadística cuya mecánica no pueda seguirse porque los números asignados al efectuar la medición no sean los apropiados. Pero una herramienta estadística no quita ni pone significado a los números que analiza. El hecho de que los números posean uno u otro significado no es un problema que pueda resolverse con la utilización de una u otra herramienta estadística, sino desde la teoría de la medida y desde el conocimiento por parte del investigador de las propiedades de las características que estudia. Por esta razón es importante conocer la problemática relacionada con las escalas de medida: el conocimiento de esta problemática puede servir, al menos, para saber si, con los números disponibles, tiene o no sentido efectuar determinado tipo de operaciones.

Programas informáticos para el análisis de datos

Hasta hace pocos años, la mayor parte de los procedimientos estadísticos se aplicaban con la ayuda de una calculadora de bolsillo. Lógicamente, los manuales de análisis de datos estaban diseñados teniendo en cuenta esta circunstancia. Afortunadamente los tiempos han cambiado y actualmente los analistas de datos disponen de ordenadores y programas informáticos capaces de efectuar los cálculos más complejos con suma rapidez y con el mínimo esfuerzo.

Este nuevo escenario ha guiado tanto la selección de los contenidos de este manual como la forma de exponerlos. Por un lado, se han incluido algunos procedimientos (particularmente en los siguientes volúmenes) cuya aplicación sería impensable (o muy difícil) sin la ayuda de un ordenador. Por otro, la exposición se ha centrado no tanto en el aparato matemático asociado a cada procedimiento como en el aprendizaje orientado a saber elegir en cada momento el procedimiento apropiado, a ejecutarlo con un programa informático y a interpretar correctamente los resultados. Por supuesto, esto no implica renunciar a las fórmulas matemáticas, pues muchas de ellas son muy útiles, no sólo para definir algunos conceptos, sino para entenderlos. Pero ya no es necesario seguir invirtiendo tiempo en hacer a mano o con una calculadora de bolsillo algunos cálculos que pueden hacerse de forma mecánica, sencilla, rápida y libre de errores con un ordenador.

La lista de programas informáticos disponibles para el análisis de datos es interminable. Muchos de ellos son *generales:* incluyen la mayoría de las técnicas estadísticas que un analista puede necesitar; otros muchos son *específicos:* se centran en una técnica concreta o en un conjunto reducido de técnicas. Entre los de carácter general (que son los que más nos interesan aquí, pues estudiaremos diferentes técnicas) destacan, entre otros, SPSS, SAS, R/S-Plus, Minitab y Stata. Y entre ellos, quizá el SPSS sea el de mayor implantación tanto en el ámbito académico como en el profesional: a su innegable potencial para el análisis hay que añadir sus prestaciones como base de datos y su facilidad de manejo. Por tanto, los ejemplos de cómo se aplican las diferentes técnicas estadísticas se basarán en el SPSS 17 (última versión disponible en la fecha de edición de este manual).

Ejercicios

1.1. A continuación se enumeran algunas de las características con las que puede encontrarse un analista de datos en su trabajo cotidiano. El objetivo de este ejercicio es el de aprender a identificar el nivel de medida (nominal, ordinal, de intervalos, de razón) que puede alcanzarse con estas características. Para responder correctamente a esta pregunta deben tenerse en cuenta las consideraciones hechas en el apartado *Rol de las escalas de medida*.

a. Percepción subjetiva del dolor.
b. Grupo de tratamiento (experimental, control).
c. Satisfacción con un determinado servicio.
d. Peso de los recién nacidos.
e. Tiempo de reacción.
f. Calidad percibida del estado de salud general.
g. Rendimiento en el test de inteligencia Raven.
h. Actitud hacia el aborto (en contra, indiferente, a favor).
i. Rendimiento en una prueba de cálculo numérico.
j. Nivel socioeconómico (bajo, medio, alto).
k Número de aciertos en una prueba de rendimiento.
l. Calidad del material recordado.
m. Nivel de ansiedad.
n. Intensidad del ruido ambiental.
ñ. Años de experiencia educativa de un profesor.
o. Color de un estímulo (rojo, amarillo, verde, azul).
p. Dosis de un fármaco (0 mg, 100 mg, 250 mg, 500 mg).
q. Grado de dificultad de una pregunta.
r. Nivel de alcohol en sangre (g/l).
s. Consumo de alcohol (nulo, bajo, medio, alto).
t. Número de cigarrillos/día.
u. Tabaquismo (fumadores, exfumadores, no fumadores).
v. Puntuaciones en la escala de depresión de Hamilton.
w Número de accidentes de tráfico ocurridos en fin de semana.
x. Tipo de ideología política (izquierda, centro, derecha).
y. Nivel de conservadurismo medido en el continuo izquierda-derecha.
z. Tipo de tratamiento antidepresivo (farmacológico, psicológico, mixto).

1.2. A continuación se describen varios estudios, de forma resumida, con el objetivo de que sean clasificados en función del nivel de indagación (descriptivo, relacional, explicativo) en el que es posible situar sus conclusiones.

a. Se ha llevado a cabo un estudio para determinar si los dibujos actúan como facilitadores o entorpecedores del aprendizaje de palabras en niños de 3 y 4 años. Se han seleccionado aleatoriamente 80 niños de una escuela infantil y, tras repartirlos en dos grupos al azar, a la mitad de ellos se les ha enseñado nuevas palabras sin utilizar ilustraciones y a la otra mitad se les ha enseñado las mismas palabras ilustradas con sencillos dibujos. Tras el entrenamiento se ha registrado el número de palabras aprendidas por cada niño.

b. Por información recogida en estudios previos se sabe que, en la población de madrileños mayores de 15 años, la proporción de fumadores, exfumadores y no fumadores es de 0,30, 0,12 y 0,58, respectivamente. Se desea averiguar si, en la población de jóvenes con edades comprendidas entre los 15 y los 25 años, se reproduce esa misma pauta. Para ello, se ha seleccionado una muestra aleatoria de 250 sujetos en la que se han encontrado 88 fumadores, 12 exfumadores y 150 no fumadores.

c. Según sugieren algunos trabajos, los niños con problemas perceptivos aumentan, con entrenamiento adecuado, su rendimiento en preguntas del test de Raven que habitualmente no resuelven por carecer de las estrategias adecuadas. Con el fin de obtener evidencia adicional sobre esto, un investigador ha seleccionado una muestra aleatoria de 10 niños con problemas perceptivos y les ha pasado el test de Raven para obtener una medida inicial en el test. Después ha entrenado a los 10 niños durante 2 meses en tareas de percepción de formas y, terminado el entrenamiento, ha vuelto a pasarles el test para obtener una nueva medida con la que comparar la medida inicial.

d. Un investigador sospecha que los hombres y las mujeres difieren en sus actitudes hacia el aborto. Para confirmar sus sospechas selecciona aleatoriamente una muestra de hombres y otra mujeres, les pasa una escala para medir su actitud y compara los resultados obtenidos.

e. Un investigador desea comprobar si la ingestión de alcohol reduce la capacidad de los sujetos para reaccionar a letras presentadas mediante taquistoscopio. Para ello, forma 10 parejas aleatorias de tal forma que los sujetos de cada pareja están igualados en agudeza visual. Un sujeto de cada pareja, seleccionado al azar, recibe una determinada dosis de alcohol. Al cabo de un tiempo preestablecido se presenta la serie de letras y se registra el número de aciertos de cada sujeto.

f. Los resultados obtenidos en algunos trabajos sugieren que el grado de deterioro de la relación entre los enfermos terminales y sus cuidadores podría deberse, entre otras causas, a la cantidad de tiempo que un cuidador permanece con el mismo enfermo. Para obtener información adicional sobre esta problemática, se han formado cuatro grupos de parejas cuidador-enfermo basados en los años de duración de la relación (0-2, 3-5, 5-8 y 8+). Se ha evaluado la calidad de la relación de cada pareja mediante cuestionarios y entrevistas y se han comparado los resultados de los cuatro grupos.

g. Diversas investigaciones ponen de manifiesto que el rendimiento está relacionado con la ansiedad de ejecución de la siguiente manera: cuando la ansiedad es baja o alta, el rendimiento es bajo; cuando la ansiedad es media, el rendimiento es alto (a este efecto se le llama ley de Yerkes-Dodson). Para contrastar este tipo de relación, se ha seleccionado una muestra aleatoria de sujetos a los que se les ha medido el nivel de ansiedad inmediatamente antes de presentarles la tarea de solución de problemas que tenían que realizar.

h. Se ha aplicado un determinado tratamiento a 50 personas (25 hombres y 25 mujeres) con síntomas claros de fobia a los perros. Se han recuperado por completo 22 hombres y 18 mujeres. El objetivo del estudio es averiguar si la eficacia del tratamiento es distinta en los hombres y en las mujeres.

i. En el contexto de la valoración que se hace anualmente de la calidad percibida de las cafeterías de la universidad, se ha estudiado la evolución que han experimentado esas valoraciones entre 2005 y 2009. La idea de los investigadores es que las valoraciones han ido mejorando con los años.

j. La Dirección General de Tráfico ha encargado a un equipo de investigación de la UAM un informe acerca del grado en que los tiempos de reacción a estímulos visuales se ven alterados por la ingestión de alcohol. Para elaborar el informe, el equipo de investigación ha diseñado un estudio con tres grupos. A los sujetos del primer grupo se les ha administrado un placebo sin alcohol; a los del segundo grupo se les ha administrado alcohol hasta conseguir un nivel en sangre de 0,25 g/l; a los del tercer grupo se les ha administrado alcohol hasta conseguir un nivel en sangre de 0,50 g/l. Todos los sujetos han sido sometidos a una prueba de discriminación para obtener un registro del tiempo de reacción medio invertido en responder a una serie de estímulos visuales.

k. Un psicólogo cree que la opinión que un paciente tiene sobre su terapeuta va cambiando a más favorable a medida que avanza el proceso terapéutico. El psicólogo tiene, incluso, algunos datos que ha ido recogiendo preguntando a algunos pacientes la opinión sobre su terapeuta en dos momentos de la terapia: recién iniciada y a punto de finalizar.

l. Algunos trabajos señalan que los trastornos de tipo depresivo afectan al 32% de las personas en paro. Un investigador sospecha que esta cifra es demasiado alta y decide obtener alguna evidencia sobre ello. Selecciona una muestra aleatoria de sujetos en paro y registra cuántos de ellos muestran trastornos de tipo depresivo.

m. En una muestra aleatoria de niños con problemas de enuresis se ha aplicado un tratamiento cognitivo-conductual y se ha registrado el número de recuperaciones.

n. Un profesor ha elaborado un examen con 10 preguntas. Antes de utilizarlo como instrumento de evaluación de su asignatura desea conocer algunas de las propiedades de las preguntas que ha elaborado. Entre esas propiedades, una que le interesa especialmente es que no todas ellas tengan un nivel de dificultad homogéneo, es decir, que haya unas preguntas más fáciles y otras más difíciles. Para obtener información sobre esta cuestión, pasa el examen a una muestra de estudiantes y registra el número de aciertos en cada pregunta.

ñ. Los resultados de algunos trabajos sugieren que el grado de deterioro de la relación entre parejas jóvenes (menores de 35 años) podría deberse, entre otras causas, a la diferencia de edad existente entre los miembros de la pareja. Para obtener información adicional sobre esta cuestión, se han formado cuatro grupos de diferencia de edad (0-3 años, 4-7 años, 8-11 años y 12-15 años). De cada grupo de edad se ha seleccionado una muestra de parejas y en cada pareja se ha evaluado la calidad de la relación existente utilizando cuestionarios y entrevistas. Finalmente se han comparado los resultados de los cuatro grupos.

o. Un investigador sospecha que las preguntas de los cuestionarios de personalidad poseen un significado especial en función del contexto definido por el cuestionario del que forman parte. Esto haría que preguntas similares fueran respondidas de forma distinta por los mismos sujetos cuando esas preguntas forman parte de cuestionarios diferentes. Para confirmar su sospecha, el psicólogo ha pasado a una muestra de sujetos tres cuestionarios de personalidad con una pregunta idéntica (tanto en la forma como en el contenido). La predicción del psicólogo era que los sujetos responderían de forma distinta a esa pregunta dependiendo del cuestionario en el que estuviera ubicada.

p. Existe la hipótesis de que los procesos de psicosis esquizofrénica van acompañados de un incremento del nivel de cobre en sangre. Esto significa que los pacientes con cuadros de psicosis esquizofrénica graves presentan un nivel de cobre en sangre más alto que los pacientes con cuadros leves. Un psicólogo clínico cree haber descubierto un tratamiento

mixto (fármaco-terapia) capaz de reducir el nivel de cobre en sangre. Para comprobar si esto es cierto elige una muestra de pacientes esquizofrénicos y mide en cada uno de ellos el nivel de cobre en sangre antes y después de aplicarles el nuevo tratamiento.

q. Con el fin de estudiar el efecto de ciertas variables motivacionales sobre el rendimiento en tareas de logro, un psicólogo ha diseñado dos programas de entrenamiento motivacional y los ha aplicado a dos grupos de sujetos seleccionados al azar. Un tercer grupo no ha recibido entrenamiento pero ha realizado la misma tarea que los sujetos entrenados (grupo control). El rendimiento de los sujetos se ha evaluado utilizando diferentes tareas (anagramas, rompecabezas, etc.).

r. Un educador está interesado en comprobar si las puntuaciones de una prueba de razonamiento abstracto se mantienen constantes o se modifican entre los 7, 8 y 9 años de edad. Para tal fin, selecciona una muestra aleatoria de niños de 7 años y les mide su nivel de razonamiento abstracto. Vuelve a efectuar el mismo registro a los 8 y a los 9 años.

s. Para estudiar el efecto de la intensidad del ruido ambiental sobre la ejecución de una tarea viso-motora compleja, un psicólogo ha seleccionado una muestra de sujetos y los ha distribuido al azar en tres grupos. Cada grupo ha sido sometido a una condición de ruido ambiental de diferente intensidad (baja, media, alta) mientras realizaban la tarea.

t. En varios trabajos clásicos sobre memoria se ha intentado estudiar el efecto del paso del tiempo presentando un determinado material a un grupo de sujetos y evaluando la calidad del recuerdo de ese material tras diferentes intervalos de tiempo. En un estudio concreto, a una muestra aleatoria de sujetos se les entregó una historia escrita para que la memorizaran durante 20 minutos. Pasado ese tiempo, se dejó transcurrir una hora y se pidió a los sujetos que escribieran en un papel la historia que habían memorizado. Un grupo de expertos evaluó la calidad del recuerdo. Transcurrido un día se volvió a pedir a los sujetos que volvieran a escribir la historia tal como la recordaban. Y lo mismo se hizo al cabo de una semana y al cabo de un mes.

u. Un investigador desea evaluar la eficacia de tres terapias diferentes para tratar problemas de ansiedad. Selecciona al azar pacientes con trastorno de ansiedad y forma, también al azar, tres grupos. Aplica a cada grupo una terapia y, tras ello, toma una medida del nivel de ansiedad de cada sujeto.

v. El departamento de ventas de una empresa ha formulado a un equipo de expertos en percepción una consulta relacionada con el impacto capaz de producir la cantidad de iluminación sobre el número de ventas. Para responder a la consulta, el grupo de expertos ha diseñado un estudio con cuatro niveles de intensidad luminosa y ha registrado el número de ventas en una muestra aleatoria de tiendas bajo los 4 niveles de intensidad luminosa (controlando la hora del día y los días de la semana).

w. Muchos trabajos sobre aprendizaje permiten concluir que el desempeño de los sujetos es tanto mejor cuanto mayor es la recompensa (refuerzo) que reciben. En un estudio con animales se han formado aleatoriamente tres grupos de ratones sedientos. Los ratones de cada grupo han sido recompensados (reforzados) con diferentes cantidades de agua (5, 10 y 15 cc) por recorrer un laberinto. Como medida del aprendizaje se ha utilizado el número de ensayos que ha necesitado cada ratón para aprender a recorrer el laberinto sin errores.

x. El director de un colegio desea saber si la experiencia educativa de los profesores tiene o no algo que ver con el rendimiento en una determinada asignatura. Para ello, ha decidido comparar los resultados obtenidos por un grupo de profesores con cinco años o menos

de experiencia con los obtenidos por un grupo de profesores con más de cinco años de experiencia.

y. El Ayuntamiento de Madrid encarga a un equipo de psicólogos el diseño de una campaña de persuasión que intente mejorar la actitud de la población madrileña hacia los enfermos de sida. Al comenzar el trabajo, el equipo de psicólogos decide obtener evidencia sobre si una técnica persuasiva basada sólo en imágenes será o no lo bastante eficaz. Para ello, registra en una muestra de personas su actitud (positiva, negativa) antes y después de una sesión de persuasión.

z. En un estudio relacionado con la problemática herencia-medio, un investigador ha conseguido reunir a 20 pares de gemelos monocigóticos (gemelos con genes idénticos) para estudiar el peso de la herencia en el cociente intelectual. Su intención es medir el cociente intelectual de los gemelos y determinar el grado de parecido existente entre ellos.

Soluciones

1.1. Las soluciones marcadas con un asterisco se refieren a variables de las que señalado que no pueden ser consideradas estrictamente ordinales. Ya hemos argumentado que estas variables pueden tratarse, a efectos prácticos, como si estuvieran medidas con una escala de intervalos:

a. Ordinal*.
b. Nominal.
c. Ordinal*.
d. Razón.
e. Razón.
f. Ordinal*.
g. Ordinal*.
h. Ordinal.
i. Ordinal*.
j. Ordinal.
k. Razón.
l. Ordinal*.
m. Ordinal*.
n. Razón.
ñ. Razón.
o. Nominal.
p. Razón.
q. Ordinal*.
r. Razón.
s. Ordinal.
t. Razón.
u. Nominal.
v. Ordinal*.
w. Razón.
x. Nominal.
y. Ordinal*.
z. Nominal.

1.2. *a*. Nivel explicativo. Hay control de las condiciones del estudio (palabras sin dibujos, palabras con dibujos) y asignación aleatoria de los sujetos a las condiciones.

b. Nivel descriptivo. El estudio se limita a observar (constatar) cómo es algo.

c. Nivel relacional. Es posible determinar si se produce una mejora, pero no si esa mejora se debe al entrenamiento (no hay forma de saber, por ejemplo, si la posible mejora es debida simplemente al paso del tiempo).

d. Nivel relacional. No hay asignación aleatoria a las condiciones del estudio (hombres, mujeres). Es posible averiguar si los hombres y las mujeres difieren en su actitud pero no el motivo de esa diferencia.

e. Nivel explicativo. Hay control de las condiciones del estudio (dosis de alcohol) y asignación aleatoria de los sujetos a las condiciones.

f. Nivel relacional. No hay asignación aleatoria de los sujetos a las condiciones del estudio. Es posible saber si los grupos formados difieren, pero no si las diferencias se deben a los años de relación o a otra cosa.

g. Nivel relacional. No hay asignación aleatoria de los sujetos a las condiciones del estudio (nivel de ansiedad). Es posible saber si los grupos formados difieren, pero no si esas diferencias se deben al nivel de ansiedad o a alguna otra variable.

h. Nivel relacional. Es posible saber si el número o porcentaje de recuperaciones es distinto en los hombres y en las mujeres, pero no el verdadero motivo de la diferencia.

i. Nivel relacional. Es posible averiguar si la calidad percibida mejora con el tiempo, pero no es posible asegurar que la mejora sea debida al paso del tiempo.

j. Nivel explicativo. Hay control de las condiciones del estudio (nivel de alcohol en sangre) y asignación aleatoria de los sujetos a las condiciones.

k. Nivel relacional. Es posible averiguar si la opinión de los pacientes cambia con la evolución de la terapia, pero no es posible asegurar que la evolución de la terapia sea la responsable de ese cambio.

l. Nivel descriptivo. El estudio se limita a observar el número o porcentaje de parados con problemas de depresión.

m. Nivel relacional. Es posible determinar si se produce un cambio pero no es posible asegurar que el cambio se deba al tratamiento (podría deberse al paso del tiempo, podría existir un efecto placebo, etc.).

n. Nivel descriptivo. Puesto que sólo se ha registrado el número de aciertos de cada pregunta, lo único que puede saberse es si existen diferencias en la proporción de aciertos..

ñ. Nivel relacional. No hay asignación aleatoria de los sujetos a las condiciones del estudio (diferencia de años). Es posible saber si los grupos formados difieren, pero no el motivo de la diferencia.

o. Nivel explicativo. Hay control sobre las condiciones del estudio (los dos contextos que rodean a la pregunta) y ambas se aplican a los mismos sujetos.

p. Nivel relacional. Es posible determinar si se produce un cambio pero no es posible asegurar que el cambio se deba al tratamiento (podría deberse al paso del tiempo, podría existir un efecto placebo, etc.).

q. Nivel explicativo. Hay control de las condiciones del estudio (tipo de entrenamiento) y asignación aleatoria de los sujetos a las condiciones.

r. Nivel relacional. Es posible determinar si el razonamiento abstracto cambia, pero no es posible asegurar que ese cambio se deba al paso del tiempo.

s. Nivel explicativo. Hay control de las condiciones del estudio (intensidad del ruido) y asignación aleatoria de los sujetos a las condiciones.

t. Nivel relacional. Es posible establecer si el recuerdo cambia con el paso del tiempo, pero no es posible asegurar que el cambio observado se deba al paso del tiempo.

u. Nivel explicativo. Hay control de las condiciones (tipo de terapia) y asignación aleatoria de los sujetos a las condiciones.

v. Nivel explicativo. Hay control de las condiciones (intensidad luminosa) y las cuatro condiciones se aplican a las mismas tiendas.

w. Nivel explicativo. Hay control de las condiciones del estudio (cantidad de recompensa) y asignación aleatoria de los ratones a las condiciones.

x. Nivel relacional. No hay control de las condiciones del estudio (la experiencia del profesor viene dada) ni asignación aleatoria de los sujetos a las condiciones.

y. Nivel relacional. Es posible determinar si se produce un cambio pero no es posible asegurar que el cambio se deba a la intervención (podría deberse al paso del tiempo, podría existir un efecto placebo, etc.).

z. Nivel relacional. No hay asignación aleatoria de los sujetos a las condiciones del estudio. Es posible determinar el grado de parecido o diferencia existente entre los cocientes intelectuales de los gemelos, pero no aislar la causa del parecido o la diferencia.

2

Conceptos previos

Ya en el capítulo anterior han empezado a aparecer algunas ideas fundamentales relacionadas con el análisis de datos (niveles de indagación, escalas de medida). En este capítulo continuaremos revisando algunos conceptos básicos (variable, población, muestra, parámetro, estadístico, muestreo) y ofreciendo una exposición resumida de la teoría de la probabilidad, la cual, según tendremos ocasión de constatar repetidamente, constituye el argumento matemático en el que se basan gran parte de los procedimientos estadísticos que estudiaremos en los próximos capítulos.

Tipos de variables

Una *variable* es la *representación numérica de una característica sometida a medición*. Recibe ese nombre porque, al medir una característica en un conjunto de elementos (por ejemplo, la *altura* en un grupo de sujetos), los valores que se obtienen no son idénticos en todos los elementos medidos (las alturas de los sujetos *varían*). Normalmente, la característica medida (la altura) también recibe el nombre de variable, aunque hay quien prefiere reservar el término para el resultado de la medición (los valores obtenidos al medir la altura).

Los niveles o escalas de medida descritos en el capítulo anterior sirven para hacer una primera clasificación de los diferentes tipos de variables. En principio, podríamos decir que existen tantos tipos de variables como escalas o niveles de medida: nominal, ordinal, de intervalos y de razón. Pero las consideraciones teóricas del capítulo anterior deben ser revisadas cuando se adopta un punto de vista práctico. Por un lado, la distinción entre medidas de intervalos y de razón carece por completo de relevancia para el análisis de datos: aunque las operaciones aritméticas que tiene sentido hacer con los números que se obtienen con esas dos medidas no son las mismas, las operaciones estadísticas sí lo son (en estadística se suele trabajar con *distancias*; y eso convierte en irrelevante el hecho de que el cero de la escala sea absoluto o no). Por otro lado, las reflexiones ya hechas en relación con la medida ordinal ponen de manifiesto que existen serios inconvenientes para asumir que todas las variables *teóricamente ordinales* son del mismo tipo. Estas consideraciones (basadas en las reflexiones hechas en el capítulo anterior sobre el *rol de las escalas de medida*) justifican, en nuestra opinión, una clasificación de los diferentes tipos de variables en sólo dos: *categóricas* y *cuantitativas*.

Llamamos variables **categóricas** a las mediciones resultantes de aplicar una escala nominal (*sexo*: hombre, mujer; *tipo de tratamiento*: A, B, control; *resultado del tratamiento*: recuperados, no recuperados; *resultado de un ensayo*: acierto, error; *tipo de metas motivacionales*: aprendizaje, ejecución; etc.). Se incluyen aquí las variables que, aun siendo ordinales, sólo tienen unas pocas categorías distintas (*clase social*: baja, media-baja, media, media-alta, alta; *nivel de estudios*: primarios, secundarios, medios, superiores; etc.).

Llamamos variables **cuantitativas** a las mediciones que resultan al aplicar una escala de intervalos o de razón (la *temperatura* medida en grados Celsius, la *altura* medida en cm, el *peso* medido en kg, el *tiempo de reacción* medido en milisegundos, el *número de aciertos* en una prueba de rendimiento, etc.). Incluimos aquí las variables que, aun no alcanzando el nivel de medida de intervalos (como ocurre, por ejemplo, con las puntuaciones en una escala de dolor percibido), no está claro que puedan reducirse a un nivel de medida estrictamente ordinal. Esta última afirmación es especialmente relevante si se tiene en cuenta que en muchas áreas de conocimiento se utilizan escalas para medir actitudes, satisfacción, habilidades, emociones, calidad de vida, estado de salud percibido, etc. Este tipo de escalas arrojan, en teoría, mediciones ordinales y, por tanto, variables también ordinales, pero de ese tipo de variables que ya hemos calificado como *no estrictamente ordinales* y, por tanto, de las que, en la práctica, pueden tratarse como si en realidad fueran cuantitativas. Podríamos decir que las herramientas estadísticas que permiten obtener información útil con estas variables que estamos calificando de *no estrictamente ordinales* son las herramientas diseñadas para analizar variables cuantitativas (de intervalos o de razón). Y no olvidemos que uno de los principales objetivos del análisis es el de extraer información útil de los datos.

Las variables *cuantitativas* pueden ser discretas o continuas. Una variable se considera **discreta** cuando entre dos valores consecutivos no puede darse un valor intermedio; éste es el caso de variables como el número de hijos o la proporción de aciertos en un test (se pueden tener 2 o 3 hijos, pero no 2,7; y la proporción de aciertos en un test toma valores discretos, aunque tenga decimales, porque procede del número de aciertos, que es una variable discreta). Una variable se considera **continua** cuando entre dos valores consecutivos siempre es posible encontrar un valor intermedio; éste es el caso de variables como la edad o los tiempos de reacción (se puede tener 21 o 22 años, pero también 21,3 o 21,34571; el número de decimales depende de la precisión que seamos capaces de conseguir). En la práctica, dado que la precisión con la que es posible medir tiene sus limitaciones y que pretender medir con una precisión ilimitada no tiene ningún sentido, todas las variables son, de hecho, discretas. No obstante, la distinción entre variables discretas y continuas tiene su importancia teórica pues, según veremos, los modelos de probabilidad diseñados para uno y otro tipo de variables tienen sus peculiaridades.

Para poder trabajar cómodamente con variables es importante estar familiarizado con la *notación* que utilizaremos. Por lo general, a las variables las representaremos con letras latinas mayúsculas: X, Y, Z . Y para distinguir una variable de los valores concretos que toma, añadiremos un subíndice: X_i, Y_i, Z_i. El subíndice no tiene nada que ver con el valor concreto que toma la variable, sino con la posición que ocupa ese valor en el conjunto de valores de la variable: X_1 se refiere al primer valor de la variable X; X_2 se refiere al segundo valor de la variable X; X_n se refiere al enésimo –el último – valor de la variable X. Así, si la variable X toma los valores 3, 7, 9, 12 y 15, entonces $X_1 = 3$, $X_2 = 7$, ..., y $X_5 = 15$. Ocasionalmente utilizaremos letras minúsculas para representar una variable (tal es el caso de las puntuaciones diferenciales); pero siempre quedará claro de qué se está hablando.

Población y muestra

El análisis de datos debe ser entendido, ante todo, como un conjunto de herramientas al servicio de la investigación empírica. Ahí es donde encaja como conjunto de procedimientos diseñados para organizar los datos, extraer información útil y elaborar conclusiones. En ocasiones, los objetivos de un estudio podrán cubrirse simplemente resumiendo la información mediante herramientas descriptivas; pero, normalmente, será necesario aplicar herramientas inferenciales para poder efectuar comparaciones y estudiar relaciones.

Ocurre, además, que, por lo general, las conclusiones de un estudio se basan en datos particulares. Si se desea, por ejemplo, probar la eficacia de un nuevo tratamiento para aliviar el insomnio, es lógico pensar que no será posible reunir a todos las personas que padecen insomnio; más bien habrá que conformarse con aplicar el tratamiento sólo a algunos pacientes; por otro lado, tampoco parece razonable aplicar a todos los pacientes con insomnio un tratamiento cuya eficacia se desconoce. Utilizar sólo unos pocos elementos del total es algo con lo que hay que lidiar casi siempre que se realiza un estudio: para conocer, por ejemplo, la opinión de los españoles sobre la eutanasia no será posible recoger la opinión de todos los españoles; para saber cómo reaccionan a un estímulo visual las personas mayores de 60 años no será posible presentar el estímulo a todos los humanos mayores de 60 años; etc. Ocasionalmente se tendrá acceso a todos los elementos que se desee estudiar; pero eso será más bien la excepción y no la regla.

Ahora bien, aunque sólo se utilicen unos pocos pacientes, o unos pocos españoles, o unos pocos humanos, lo habitual es que las conclusiones de un estudio no queden restringidas a esos pocos sujetos. Lo que realmente suele interesar es poder utilizar la información disponible para elaborar conclusiones sobre el conjunto total de sujetos de la misma clase (todos los pacientes con insomnio, todos los españoles mayores de 60 años, todos los humanos). A este salto de lo particular a lo general es a lo que llamamos *inferencia estadística*.

La inferencia estadística exige utilizar, por un lado, procedimientos que ayuden a efectuar correctamente el salto de lo particular a lo general y, por otro, procedimientos que garanticen que ese salto se apoya en una buena base. Tan importante es disponer de una buena técnica de análisis de datos para realizar la inferencia como *seleccionar* apropiadamente los datos que se van a analizar para, de esta manera, proporcionar una buena base de apoyo a la inferencia. Las *técnicas de muestreo* se encargan de garantizar que la inferencia se apoya en una buena base. Y las herramientas estadísticas englobadas bajo la denominación general de *análisis de datos* se encargan de garantizar que la inferencia se desarrolla correctamente. De esto último tratan los próximos capítulos, pero antes conviene repasar algunos conceptos fundamentales que ayudarán a entender lo demás.

Una **población** o *universo* es un *conjunto de elementos (sujetos, objetos, entidades abstractas, etc.) que poseen una o más características en común*. En general, el término población hace referencia al conjunto total de elementos que interesa estudiar y queda definida cuando se hacen explícitas las características que esos elementos comparten. Ejemplos de poblaciones son: las personas empadronadas en una comunidad autónoma, todos los hombres mayores de 30 años, los pacientes que sufren depresión, las posibles respuestas que un sujeto podría emitir en una escala de satisfacción, el censo de votantes en unas elecciones, los números múltiplos de 3; etc.

Las poblaciones pueden ser de muy diversa índole; algunas son incluso *ficticias*, en el sentido de que, aun estando formadas por elementos observables, no todos ellos resultan acce-

sibles. Si se quiere trabajar, por ejemplo, con la población de "hombres españoles mayores de 30 años", puede ocurrir que muchos de ellos no estén censados, a otros no habrá forma de localizarlos, otros no estarán dispuestos a participar en el estudio, etc. En estas circunstancias, la población real no será exactamente la de los hombres españoles mayores de 30 años, sino otra parecida: la de los "hombres españoles mayores de 30 años a los que se ha tenido acceso". Es muy importante intentar definir con la mayor precisión posible la población con la que se va a trabajar, pues va a constituir el marco desde el que se va a iniciar la recogida de datos y sobre el que van a recaer las conclusiones del análisis.

Dependiendo del número de elementos de que constan, unas poblaciones son *finitas* y otras *infinitas*. Los pacientes que padecen depresión o los votantes censados son ejemplos de poblaciones finitas. Los números múltiplos de 3 o las posibles respuestas (tiempos de reacción) que un sujeto puede emitir en una tarea de discriminación visual son ejemplos de poblaciones infinitas. Normalmente, las poblaciones con las que interesa trabajar en las ciencias sociales y de la salud son finitas, pero tan grandes que a todos los efectos pueden considerarse infinitas. Es precisamente el hecho de que las poblaciones, por lo general, sean infinitas o estén formadas por un gran número de elementos lo que hace que la descripción exacta de sus propiedades sea un objetivo prácticamente inaccesible. Por esta razón, lo habitual es trabajar con muestras.

Una **muestra** es un *subconjunto de elementos de una población*. A diferencia de las poblaciones, que suelen ser conjuntos de elementos de gran tamaño, las muestras suelen ser conjuntos de elementos de tamaño reducido. Por supuesto, para poder describir con exactitud las propiedades de una población cualquiera, sería necesario examinar todos y cada uno de los elementos que componen esa población. Pero, dado que las poblaciones que habitualmente interesa estudiar son tan grandes que, normalmente, resulta muy difícil (si no imposible) tener acceso a todos sus elementos, son las muestras las que proporcionan la información necesaria para poder describir las propiedades de las poblaciones objeto de estudio.

El conocimiento que se va generando en la vida cotidiana acerca del mundo está, muy frecuentemente, basado en muestras: con sólo comer una vez en un restaurante nos formamos una opinión acerca de la calidad de su cocina y de su servicio; con sólo conocer a un par de personas de un determinado colectivo nos formamos una idea sobre el tipo de personas que forman ese colectivo; con sólo ver dos películas de un director de cine nos aventuramos a encasillarlo en un estilo; etc. Con el análisis de datos se hace algo parecido: se extraen conclusiones referidas a todos los elementos (población) a partir de la observación de sólo unos pocos elementos (muestra).

Ahora bien, para que estas conclusiones sean válidas es necesario que la muestra utilizada sea *representativa* de la población a la que se supone que representa, lo cual se consigue mediante las técnicas de *muestreo* (ver más adelante, en este mismo capítulo). Al hablar de los diferentes tipos de muestreo volveremos sobre el concepto de muestra y ello nos permitirá seguir profundizando en su significado.

Parámetros y estadísticos

Un **parámetro** es un *valor numérico que describe una característica poblacional*. Ya se ha definido una población como un conjunto de elementos que poseen una o más características en común. Pero los elementos de una población poseen, además, otras muchas características que no comparten o en las que no coinciden. Por ejemplo, la población de hombres españoles

mayores de 30 años está formada por elementos que tienen en común *ser hombres*, *españoles* y *mayores de 30 años*, pero en esa población es posible considerar otras muchas características en las que no todos los elementos poblacionales coinciden: el estado civil, el nivel educativo, el peso, la altura, la presión arterial, la actitud hacia la eutanasia, el estado de salud percibido, etc. Al medir, por ejemplo, el *estado de salud percibido*, se obtendrán tantos valores numéricos como elementos formen parte de la población (suponiendo que se tenga acceso a todos los elementos). Si ahora se calcula el promedio (un solo número) de esos valores numéricos se habrá definido un parámetro, pues se habrá descrito numéricamente una característica de la población: el *estado de salud percibido medio* de los hombres españoles mayores de 30 años.

En la población de personas que padecen trastorno depresivo, todos los elementos de la población coinciden en una característica específica: *padecer trastorno depresivo*. Pero existen, obviamente, otras características en las que no todos los elementos coinciden. Por ejemplo, unos pacientes serán hombres y otros mujeres. Si se tuviera acceso a todos los elementos de esa población, se podría contar el número de pacientes que son hombres (o mujeres) y eso permitiría definir un parámetro; es decir, permitiría describir numéricamente una característica de la población: *la proporción de hombres (o mujeres)* en la población de pacientes con trastorno depresivo. Así pues, existen valores numéricos como la media o la proporción (además de otros muchos que tendremos ocasión de estudiar), que cuando se refieren a alguna característica poblacional reciben el nombre de parámetros.

Hay algunas características de los parámetros que interesa resaltar. En primer lugar, los parámetros son, en general, valores poblacionales *desconocidos*: puesto que las poblaciones con las que se suele trabajar son tan grandes que sus elementos raramente resultan accesibles en su totalidad, no es posible calcular un valor numérico basado en todos los elementos. En segundo lugar, los parámetros son valores numéricos *constantes* en el sentido de que son valores únicos (es decir, no son variables): definida una población cualquiera y un parámetro en ella, ese parámetro sólo puede tomar un valor numérico concreto: en un momento dado, la *proporción de hombres* en la población de pacientes con trastorno depresivo es un valor único. Por último, es necesario señalar que para referirnos a los parámetros utilizaremos (así es como suele hacerse) *letras griegas minúsculas*: μ, σ, π, ρ, β, etc.

Un **estadístico** es un valor numérico que describe una característica muestral. Por tanto, un estadístico es a la muestra lo que un parámetro a la población. Acabamos de ver que en una población cualquiera, además de las características que la definen y que son comunes a todos los elementos, es posible definir otras muchas características en las que no todos los elementos coinciden. De una muestra, lógicamente, cabe decir lo mismo. Y una vez definida una o más de esas características en las que no todos los elementos coinciden, es posible obtener un valor numérico que las describa: a ese valor numérico se le llama *estadístico*.

De la población de hombres españoles mayores de 30 años se puede extraer una muestra de n sujetos. En esa muestra se puede definir y medir, por ejemplo, la altura. Hecho esto, es posible realizar diferentes transformaciones con los valores obtenidos: sumarlos, multiplicarlos, sumarlos y dividirlos por el número de valores, etc. Cada una de estas transformaciones es un valor numérico que describe un aspecto diferente de la característica medida (la altura). Es decir, cada una de estas transformaciones es un estadístico. Pero no todos los estadísticos poseen la misma utilidad. De hecho, muchos de ellos no tienen ninguna utilidad porque no tienen ningún significado. Otros muchos, como la *media*, la *mediana*, la *desviación típica*, la *proporción*, etc., tienen un significado y utilidad contrastados, y por esta razón se utilizan para analizar datos.

Recordemos que los *parámetros* son valores poblacionales generalmente *desconocidos* porque corresponden a elementos a los que no se tiene acceso en su totalidad. Esto sería un verdadero problema si no fuera porque cada parámetro poblacional posee su réplica muestral en un estadístico concreto susceptible de ser calculado. Esto significa que los *estadísticos* muestrales se van a utilizar para intentar formarnos una idea sobre los verdaderos valores de sus correspondientes parámetros poblacionales desconocidos. Este proceso consistente en atribuir a un parámetro el valor que toma su correspondiente estadístico se conoce con el nombre de *estimación*. La estimación es un concepto especialmente importante en estadística inferencial (y, por tanto, también en el análisis de datos); a ella dedicaremos un capítulo completo, pero antes debemos seguir profundizando en el concepto de estadístico.

Es evidente que de una población cualquiera es posible extraer más de una muestra diferente del mismo tamaño. Esto significa que, definido un estadístico, cualquiera que éste sea, su valor exacto dependerá de los valores concretos que tomen cada uno de los elementos que formen parte de la muestra obtenida. Ahora bien, de una población de tamaño[1] N es posible extraer N^n muestras diferentes[2] de tamaño n. Si en cada una de esas N^n muestras calculamos un estadístico, encontraremos que el valor de ese estadístico no siempre es el mismo; es decir, encontraremos que el valor del estadístico *varía* de una muestra a otra. Esto significa que un estadístico no es un valor numérico constante (como lo es un parámetro), sino que es una *variable*: su valor concreto *varía* dependiendo de la muestra en la que se calcula.

Resumiendo, mientras un parámetro es un valor poblacional, un estadístico es un valor muestral; mientras un parámetro es, por lo general, un valor desconocido, un estadístico es un valor conocido o susceptible de ser conocido; mientras un parámetro es un valor numérico constante, un estadístico es una variable. Estas diferencias también se reflejan en la notación habitualmente utilizada para representar a unos y a otros. Mientras que los parámetros se suelen representar con letras griegas minúsculas (μ, σ, π, ρ, β, etc.), los estadísticos se suelen representar con letras latinas mayúsculas ($\bar{X}$, S, P, R, B, etc.).

Muestreo

Ya hemos señalado que uno de los objetivos fundamentales del análisis de datos es el de extraer conclusiones de tipo general a partir de unos pocos datos particulares. También hemos señalado que esto exige utilizar, por un lado, procedimientos que ayuden a efectuar correctamente ese salto (inferencia) de lo particular a lo general y, por otro, procedimientos que garanticen que el salto se apoya en una buena base. Tan importante como disponer de una buena técnica para analizar los datos es seleccionar apropiadamente los datos que se van a analizar. Qué datos se analizan condiciona la utilidad del cómo se analizan.

---

[1] Obviamente, si se está utilizando N para representar el *tamaño* de una población es porque esa población es *finita*. En una población infinita también es infinito el número de muestras distintas de tamaño n que es posible extraer.

[2] El muestreo aleatorio puede realizarse de dos maneras distintas: (1) *con reposición*, es decir, devolviendo cada elemento a la población una vez que ha sido seleccionado (lo que implica que ese elemento puede aparecer más de una vez en la misma muestra) y (2) *sin reposición*, es decir, sin devolver a la población los elementos que van siendo seleccionados. Si la muestra se obtiene con reposición, el número de muestras que es posible obtener viene dado por N^n, es decir, por las variaciones con repetición de N elementos (tamaño de la población) tomados de n en n (tamaño de la muestra). Si la muestra se obtiene sin reposición, el número de muestras posibles viene dado por $N!/(N-n)!$, es decir, por las variaciones sin repetición de N elementos tomados de n en n.

Wonnacott y Wonnacott (1990, pág. 4) recogen un ejemplo que resulta especialmente útil para ilustrar esta idea. En 1936, los editores de *Literary Digest* intentaron pronosticar el resultado de las elecciones presidenciales de Estados Unidos utilizando una muestra formada por votantes seleccionados de las guías telefónicas y de los listados de miembros de varios clubes. La muestra así obtenida presentaba (como pudo constatarse después) un fuerte sesgo hacia el bando republicano, lo cual se vio agravado, muy probablemente, por el hecho de que, de todos los cuestionarios enviados, sólo fueron contestados una cuarta parte. La muestra resultó ser tan sesgada (es decir, tan poco representativa de la población de votantes) que llevó al erróneo pronóstico de que se produciría una victoria republicana. El día de la votación se produjo la sorpresa: los republicanos obtuvieron menos del 40% de los votos y el candidato demócrata, Roosevelt, fue reelegido presidente por una aplastante mayoría. Es probable que el candidato republicano, Alf Landon (quien seguramente se había levantado esa mañana esperando ser nombrado presidente), dejara de confiar en las predicciones elaboradas a partir de encuestas basadas en muestras.

La más importante lección que debe aprenderse del error cometido por los editores de *Literary Digest* es que, cuando se intenta extraer conclusiones sobre las propiedades de una población a partir de la información contenida en una muestra de esa población, es necesario, ante todo, utilizar muestras representativas del total de la población. El no trabajar con muestras apropiadas llevará inevitablemente a que nuestras predicciones estén, ya desde el principio, condenadas al fracaso (lo que puede constituir un verdadero problema cuando, como es frecuente, esas predicciones están en la base de decisiones importantes). Por tanto, para que una muestra pueda ofrecer información satisfactoria sobre las propiedades de una población es necesario, antes que nada, que sea *representativa* de la población. Y esto sólo se consigue si todos los elementos poblacionales han tenido la oportunidad de ser elegidos.

El término **muestreo** se refiere al *proceso seguido para extraer una muestra de una población*. El muestreo puede ser de dos tipos: probabilístico y no-probabilístico. En el muestreo *probabilístico* se conoce (o puede calcularse) la probabilidad asociada a cada una de las muestras que es posible extraer de una determinada población; y cada elemento poblacional tiene asociada una probabilidad conocida (o calculable) de pertenecer a la muestra. En el muestreo *no-probabilístico* se desconoce o no se tiene en cuenta la probabilidad asociada a cada posible resultado muestral: el investigador selecciona aquella muestra que más representativa le parece o, simplemente, aquella que considera que puede extraer con mayor comodidad o menor coste (voluntarios que responden a un anuncio, alumnos matriculados en un curso o en un centro, clientes que compran un producto, pacientes que acuden a un centro de salud o a un hospital, etc.).

Lógicamente, sólo con el muestreo probabilístico se conoce la probabilidad asociada a cada resultado muestral y, consecuentemente, sólo él permite formarse una idea sobre el grado de representatividad de una muestra. Por tanto, sólo el muestreo probabilístico ofrece una base adecuada para inducir las propiedades de una población a partir de la información muestral. Esto no significa que el muestreo no probabilístico no pueda generar muestras representativas; lo que ocurre es que al utilizar un muestreo de tipo no probabilístico no se tiene información acerca de si la muestra es o no representativa. En consecuencia, ya desde ahora, se dejará a un lado el muestreo no probabilístico y se considerará en todo momento que los datos de que se dispone constituyen una muestra aleatoriamente seleccionada de su respectiva población, es decir, una **muestra aleatoria**.

En el muestreo aleatorio (selección al azar) se verifican dos importantes propiedades. En primer lugar, todos los elementos poblacionales tienen la misma probabilidad de ser elegidos;

por tanto, cualquiera de ellos puede ser elegido y ésta es una condición necesaria para obtener una muestra representativa. En segundo lugar, el resultado de cada extracción no afecta ni depende del resultado de cualquier otra; es decir, las extracciones son independientes entre sí; y ésta, según tendremos ocasión de comprobar, es una condición que asume la mayoría de los procedimientos estadísticos que estudiaremos (para profundizar en estos conceptos, puede consultarse Pardo y San Martín, 1998, págs. 45-55).

Debe tenerse en cuenta que, puesto que las poblaciones con las que se suele trabajar son desconocidas, nunca hay forma de saber si la muestra elegida es o no representativa de la población muestreada. Lo que sí se sabe es si se ha utilizado o no un método de selección que garantiza que la muestra elegida es una muestra representativa de la población. Y ese método de selección es el muestreo aleatorio.

Ahora bien, aunque el muestreo aleatorio permite obtener una muestra apropiada en la mayor parte de los contextos, en ocasiones es posible que surja la necesidad de trabajar con poblaciones cuyas características estén aconsejando alguna variante. No es éste el lugar para describir con detalle los diferentes tipos de muestreo aleatorio, pero sí nos parece conveniente ofrecer una breve descripción de los más utilizados.

En el **muestreo aleatorio *sistemático*** se comienza elaborando una lista con los N elementos poblacionales numerados de 1 a N. A continuación se fija el tamaño de la muestra que se desea obtener (n) y se efectúa una extracción al azar entre los $k = N/n$ primeros elementos (si k no es un número entero, se redondea al entero más próximo). El resto de los $n - 1$ elementos que configurarán la muestra se obtienen a partir de k. Llamando i a la posición ocupada por el primer elemento extraído, la muestra estará formada por los elementos poblacionales que ocupen las posiciones i, $i+k$, $i+2k$, $i+3k$, ..., $i+(n-1)k$.

Así, para extraer una muestra aleatoria de tamaño 100 de una población de 2.000 elementos, se comienza elaborando una lista asignando a cada elemento un número de 1 a 2.000. La constante que se debe utilizar es $k = N/n = 2.000/100 = 20$. Después, se selecciona al azar un elemento entre los 20 primeros. Si, por ejemplo, el elemento seleccionado es el que ocupa la posición $i = 9$, el resto de los elementos de la muestra serán los que ocupen en la lista las posiciones 29, 49, 69, 89, ..., 1949, 1969, 1989. Este tipo de muestreo es útil cuando se dispone de un listado de toda la población y se desea obtener una muestra aleatoria homogéneamente repartida a lo largo de toda la lista.

El **muestreo aleatorio *estratificado*** se utiliza cuando una población está formada por diferentes subpoblaciones o *estratos*. Por ejemplo, en la población de hombres españoles mayores de 30 años se pueden definir diferentes estratos según el nivel socioeconómico, el tipo de profesión, el nivel de estudios, el estado civil, etc. Con el muestreo aleatorio simple existe la posibilidad de que alguno de los estratos no esté suficientemente representado (particularmente si existen estratos muy pequeños). El muestreo aleatorio estratificado es útil cuando existe especial interés en que todos los estratos de la población tengan una adecuada representación. Se comienza definiendo los estratos e identificando los elementos que pertenecen a cada estrato. Se tienen así k estratos con tamaños N_1, N_2, ..., N_k ($N_1 + N_2 + \cdots + N_k = N$). A continuación se elaboran k listas (una por estrato) con los elementos de cada estrato debidamente numerados y se procede a extraer aleatoriamente una muestra de cada estrato mediante muestreo aleatorio simple o mediante muestreo aleatorio sistemático. La muestra total estará formada por las k submuestras extraídas.

El tamaño de las submuestras puede o no ser proporcional al tamaño de los estratos. En la *afijación simple* se asigna a todas las submuestras el mismo tamaño. En la *afijación proporcional* el tamaño de las submuestras se fija de forma proporcional al tamaño de los estratos.

Y si la variabilidad de los estratos es muy distinta, conviene extraer submuestras más grandes de los estratos con mayor varianza: *afijación óptima*. Por ejemplo, si al extraer una muestra aleatoria de tamaño 100 de una población formada por 20.000 personas con un 40% de hombres y un 60% de mujeres, queremos que esas proporciones poblacionales se mantengan en la muestra (afijación proporcional), debemos formar dos estratos (es decir, dos grupos: uno con los hombres y otro con las mujeres) y seleccionar aleatoriamente 40 sujetos del primer estrato y 60 del segundo. Si se conocieran las varianzas de la variable estudiada y la del grupo de hombres fuera muy diferente de la del grupo de mujeres, convendría seleccionar más sujetos del estrato con mayor varianza.

En el **muestreo por *conglomerados***, las unidades muestrales no son elementos individuales, sino grupos de elementos llamados *conglomerados*. En lugar de considerar que la población está formada por N elementos, se considera que está formada por k conjuntos o conglomerados de elementos. Se selecciona aleatoriamente uno o varios de esos conglomerados y se acepta como muestra el conjunto de *todos* los elementos que forman parte de ese o esos conglomerados seleccionados. Por ejemplo, en un estudio sobre desarrollo cognitivo en el que la población de referencia es la de todos los alumnos de Educación Primaria de la Comunidad de Madrid, en lugar de seleccionar una muestra aleatoria de un listado de todos los alumnos de Educación Primaria, se podrían seleccionar unos pocos colegios de la población de colegios y utilizar como muestra a todos los alumnos de los colegios seleccionados. Las ventajas de este tipo de muestreo son evidentes cuando se trabaja con poblaciones muy grandes: no se necesita un listado de todos los elementos de la población, sino sólo de aquellos que forman parte de los conglomerados seleccionados.

En el muestreo aleatorio por conglomerados puede procederse por etapas; se habla entonces de **muestreo *polietápico***. En la primera etapa se divide la población en k conglomerados y se elige uno o varios de ellos (unidades muestrales primarias); en la segunda, los conglomerados elegidos se dividen en conglomerados más pequeños y se vuelve a elegir uno o varios de ellos (unidades muestrales secundarias); etc. La muestra definitiva la componen todos los elementos de los conglomerados seleccionados en la última etapa. Obviamente, cuando se procede por etapas sólo es necesario un listado de los elementos que forman parte de los conglomerados seleccionados en la última etapa. Si, en el estudio sobre desarrollo cognitivo, la población de referencia fuese la de todos los alumnos españoles de enseñanza primaria, se podría comenzar seleccionando unas pocas comunidades autónomas; después, una provincia de cada comunidad autónoma seleccionada; después, un pueblo o ciudad de esas provincias; por último, un colegio de cada pueblo o ciudad seleccionados. Al proceder por etapas, en cada etapa y dependiendo de las características de los conglomerados que finalmente se vayan a muestrear, es posible utilizar cualquiera de los restantes métodos de muestreo aleatorio: simple, sistemático o estratificado.

Variables aleatorias

El concepto de variable[3] como *representación numérica de una característica sometida a medición* ya se ha presentado al hablar de los distintos tipos de variables (ver, en este mismo

---

[3] En este apartado se ofrece una explicación más bien intuitiva y poco formal del concepto de variable aleatoria y de sus características. Este tipo de explicación es la que nos ha parecido más apropiada para quienes se inician en el análisis de datos. El lector interesado en una exposición más formal puede consultar Amón (1984, Capítulos 3 a 6).

capítulo, el apartado sobre *Tipos de variables*). En ese momento se destacó el hecho de que una variable es la representación de una característica (sexo, altura, etc.) que no siempre que se mide toma los mismos valores, es decir, la representación de una característica que *varía*. Ha llegado el momento de señalar otra importante peculiaridad de las variables que suele interesar analizar: la *aleatoriedad* resultante del muestreo.

Una variable aleatoria[4] es una colección de números (al menos dos). En sentido estricto, hasta que no hay números, no hay variable. Pero ya sabemos que no todos los números que se asignan en el proceso de medición tienen el mismo significado, lo cual nos ha llevado a clasificar las variables como categóricas y cuantitativas.

Al medir una variable en una muestra de tamaño *n* se obtienen *n* valores. Si la variable es categórica (por ejemplo, *sexo*), los posibles valores distintos serán pocos (hombre, mujer) y cada uno de ellos se repetirá varias veces (pues todos los resultados serán hombre o mujer). Por el contrario, si la variable es cuantitativa (por ejemplo, *altura*), habrá muy pocas repeticiones o ninguna (si la medida se hace con suficiente precisión, habrá muchos valores distintos y muy pocas repeticiones de un mismo valor). Tras asignar números a los resultados del muestreo (por ejemplo, unos a los hombres y doses a las mujeres; centímetros a las alturas), en ambos casos tendremos *variables aleatorias* porque en ambos casos tendremos *números resultantes del muestreo aleatorio*.

Ahora bien, saber que la variable sexo toma unos y doses no aporta información útil (ya se sabe que la variable sexo toma unos y doses, y que eso no depende del muestreo). Lo interesante es saber cuántos unos y cuántos doses aparecen en una muestra. Es en ese momento, es decir, cuando a las categorías de la variable sexo se le asocian los resultados del muestreo, cuando se tiene una variable aleatoria. Pero centrar la atención en cuántos hombres (o mujeres) aparecen en una muestra es centrar la atención, no en la variable *sexo*, sino en una nueva variable: el *número de hombres*, que es una *variable* porque depende de la muestra concreta en la que se calcula (es decir, porque varía de muestra a muestra) y, además, es *aleatoria* porque los valores que toma son resultado del muestreo aleatorio. Por supuesto, la variable *sexo* (categórica) es estadísticamente interesante: permite formar grupos y, aunque ya se sabe qué valores toma, siempre resulta posible aplicar herramientas descriptivas para conocer con qué frecuencia toma cada valor. Pero la variable *número de hombres* (cuantitativa) es mucho más interesante: permite, según veremos, efectuar comparaciones y estudiar relaciones tomando como referencia algunos modelos teóricos de probabilidad.

Con una variable cuantitativa como la *altura* ocurre algo parecido. Aunque los valores que toma la variable tienen interés en sí mismos (pueden ser más altos o más bajos, muy parecidos entre sí o muy distintos, etc.), el hecho de que haya muchos valores distintos hace difícil formarse una idea de las características de la variable si no se utiliza algún tipo de resumen como, por ejemplo, la *altura media*. Estos resúmenes son, obviamente, cuantitativos, varían de muestra a muestra (es decir, son variables) y sus valores dependen del muestreo (es decir, son variables aleatorias); y, lo que es más interesante, permiten, según veremos, efectuar comparaciones y estudiar relaciones[5].

---

[4] Una variable aleatoria es una *función que asigna un número real, y sólo uno, a cada uno de los sucesos elementales de un espacio muestral* (el lector poco familiarizado con la teoría de la probabilidad puede revisar el apartado sobre conceptos básicos de probabilidad que se ofrece en el apéndice de este mismo capítulo).

[5] Las variables categóricas no suelen ser el objetivo primordial del análisis de datos. Esto no quiere decir que variables como el sexo, el tipo de tratamiento, o el nivel educativo no tengan interés analítico, sino que el interés del análisis suele dirigirse, no exactamente a esas variables (cuyos valores suelen ser fijos y conocidos), sino al número de veces que aparece cada uno

Centro, dispersión y forma de la distribución

De lo estudiado hasta aquí cabe deducir que el *análisis de datos* es, sobre todo, *análisis de variables aleatorias*, es decir, análisis de los números que se asignan a los resultados del muestreo aleatorio. Pero, ¿qué puede hacerse con estas variables (con estos números)?

Según veremos a lo largo de este manual (y de los siguientes volúmenes), el análisis de datos suele centrarse en la aplicación de herramientas inferenciales con el objetivo de efectuar comparaciones y estudiar relaciones. Pero, antes de eso, lo primero que suele hacerse (y que conviene hacer) con un conjunto de datos es formarse una idea lo más exacta posible acerca de las características de cada variable individualmente considerada. Y esto se consigue aplicando herramientas descriptivas. Para esto, tanto las variables aleatorias directamente resultantes del muestreo (sexo, altura) como las transformaciones que normalmente interesa hacer de ellas (número de hombres o de mujeres, altura media) deben caracterizarse prestando atención a tres propiedades fundamentales: *centro*, *dispersión* y *forma*.

1. El **centro** de una variable es el valor que más se repite (variables categóricas) o el promedio del conjunto de valores (variables cuantitativas). Indica qué valor de la variable, de todos los posibles, cabe esperar encontrar con mayor probabilidad. Puede calcularse de diferentes maneras (ver los dos siguientes capítulos), pero el más utilizado se conoce como *valor esperado* o *esperanza matemática*.

   En una muestra concreta, el valor esperado de una variable es su media aritmética. Pero una muestra concreta no es más que una de las muchas (¿infinitas?) que es posible extraer de una determinada población. El concepto de valor esperado incorpora la idea del centro que cabría esperar encontrar a la larga, es decir, del que cabría encontrar en el conjunto de todas las muestras de tamaño *n* que podrían extraerse de una determinada población; lo cual no es otra cosa que el centro (media aritmética) de la población. Y, según tendremos ocasión de comprobar más adelante, el concepto de valor esperado cobra especial relevancia cuando se utiliza para identificar el centro de muchas de las distribuciones teóricas de probabilidad (binomial, normal, etc.) que se utilizan en estadística para entender mejor el comportamiento de los datos.

2. La **dispersión** de una variable se refiere al grado de concentración o alejamiento de los valores en torno al centro de la variable. Al igual que ocurre con el centro, la dispersión de una variable puede calcularse utilizando diferentes métodos (ver los siguientes dos capítulos), pero quizá el más utilizado es la *desviación típica* (y su cuadrado, la *varianza*), que viene a ser una especie de promedio de distancias al centro de la variable.

3. La **forma** de la distribución refleja la frecuencia con la que se repite cada valor (variables categóricas) o cada rango de valores (variables cuantitativas).

   Aquí es importante distinguir entre distribuciones empíricas y distribuciones teóricas. Una distribución empírica indica cómo se distribuyen, de hecho, los valores de una variable. Una distribución teórica es una fórmula matemática (un modelo) que se utiliza para facilitar el trabajo con variables aleatorias (en realidad, las distribuciones teóricas son una de las herramientas estadísticas más útiles para un analista de datos).

---

de sus valores en una muestra concreta. Por tanto, el análisis de datos es, básicamente, análisis de datos *cuantitativos*. Cuando se habla de análisis de *datos categóricos* o de *variables categóricas* se está hablando, generalmente, del análisis de las frecuencias (datos cuantitativos) asociadas a las categorías de las variables categóricas.

Una distribución empírica está formada por los valores que toma una variable en una muestra concreta y por las frecuencias relativas asociadas a cada valor. Imaginemos que en una determinada población definimos la variable *padecer trastorno depresivo*, con posibles valores "sí" y "no"; extraemos al azar una muestra de esa población y asignamos un 1 a las personas que padecen depresión y un 0 a las que no la padecen; tendremos, por un lado, una variable aleatoria (unos y ceros resultantes del muestreo) y, además, el número o proporción de unos y ceros; es decir, tendremos la *distribución empírica* formada por los valores que toma la variable (unos y ceros) y por las frecuencias relativas asociadas a cada valor (proporciones de unos y ceros). Imaginemos ahora que el 10% de las personas de la población padece depresión; en este nuevo escenario es posible utilizar el cálculo de probabilidades (en concreto, una *distribución teórica* llamada binomial; ver Capítulo 3) para conocer la probabilidad asociada a cada posible resultado muestral.

Otro ejemplo. Imaginemos que seleccionamos una muestra al azar de una determinada población y medimos la *altura* de los sujetos; los números (por ejemplo, centímetros) resultantes del muestreo constituyen una variable aleatoria; asociando a esos números la frecuencia relativa con la que aparecen tendremos la *distribución empírica* de la variable altura. Imaginemos ahora que asumimos que, en la población muestreada, las alturas de los sujetos se distribuyen en forma de campana (muchos casos en torno al centro y pocos en las orillas); es decir, imaginemos que las alturas de los sujetos se parecen a una *distribución teórica* llamada *normal* (ver Capítulo 5). En este nuevo escenario es posible utilizar la distribución teórica normal para conocer la probabilidad asociada a cada posible resultado muestral.

A la combinación formada por los valores de una variable aleatoria y por las probabilidades asociadas a cada uno de esos valores se le suele llamar *función de probabilidad* o *distribución de probabilidad*. Aquí, con frecuencia, también nos referiremos a esta combinación simplemente como *distribución*[6], intentando dejar claro en cada caso si se trata de una distribución empírica o teórica.

Así pues, para formarnos una idea apropiada de las características de una variable aleatoria vamos a prestar atención a tres propiedades: centro, dispersión y forma de la distribución. El centro es una especie de representante del resto de valores; indica en torno a qué valor es más probable encontrar casos. La dispersión ayuda a precisar si el centro es o no un buen representante del resto de valores (según veremos, desempeña un papel esencial en la inferencia estadística). La forma de la distribución, por último, permite detectar dónde tienden a agruparse los valores y si existen valores que se alejan llamativamente de los demás; y, lo que es más importante, cuál es la probabilidad asociada a cada valor de la variable y, consecuentemente, cuál es la probabilidad asociada a cada posible resultado muestral.

---

[6] En este contexto es importante recordar la distinción ya establecida entre variables *discretas* (entre dos valores consecutivos no puede darse un valor intermedio; por ejemplo, el *número de aciertos*) y *continuas* (entre dos valores consecutivos siempre es posible encontrar un valor intermedio si se mide con suficiente precisión; por ejemplo, la *edad*). Esta distinción entre variables lleva asociada una distinción entre distribuciones de probabilidad que gusta mucho enfatizar a los estadísticos. En una distribución discreta, cada valor de la variable tiene asociada una probabilidad concreta (por ejemplo, la probabilidad de obtener tres caras en cinco lanzamientos de una moneda, o la probabilidad de padecer trastorno depresivo). En una distribución continua no existe tal cosa; la probabilidad asociada a un valor concreto es nula (si se define una altura con muchos decimales, la probabilidad de que un sujeto tenga exactamente esa altura es nula; de hecho, en las distribuciones continuas se habla de *densidad* en lugar de probabilidad). Esto puede entenderse fácilmente si se tiene en cuenta que la probabilidad del conjunto de posibles valores de una variable vale 1 y que esa probabilidad hay que repartirla entre los teóricamente infinitos valores de la variable continua.

Probabilidad

La teoría de la probabilidad es el aparato matemático en el que se basa la estadística para mejorar la descripción de los datos y, sobre todo, para hacer inferencias de lo particular (muestra) a lo general (población). Entender correctamente muchos de los procedimientos estadísticos que estudiaremos (al menos, algunos aspectos concretos de esos procedimientos) requiere estar familiarizado con algunos conceptos básicos de la teoría de la probabilidad.

Por supuesto, este apartado no es, ni mucho menos, un curso sobre teoría de la probabilidad (para ello puede recurrirse a cualquiera de los excelentes manuales de probabilidad existentes en el mercado). Este apartado incluye únicamente los conceptos de probabilidad que es necesario manejar para poder trabajar con las distribuciones de probabilidad que se estudian más adelante.

Lo que se hace al analizar datos es extraer muestras aleatorias y calcular números con distribución de probabilidad conocida para poder interpretar mejor esos números y para poder tomar decisiones a partir de ellos. Por tanto, lo que conviene saber de la teoría de la probabilidad es, básicamente, la parte relacionada con la selección de muestras aleatorias y con las distribuciones de probabilidad asociadas a los valores muestrales (números) que se calculan en ellas.

Espacio muestral y sucesos

Llamamos *experimento aleatorio* a cualquier acción cuyo resultado no puede predecirse con certeza. Lanzar una moneda al aire y observar el resultado (no podemos predecir con certeza si saldrá cara o cruz) o medir la altura de un sujeto elegido al azar (no podemos predecir con certeza cuál será su altura exacta) son experimentos aleatorios.

El **espacio muestral** (E) es el conjunto de posibles resultados de un experimento aleatorio. En el experimento aleatorio consistente en lanzar una moneda y observar el resultado, el espacio muestral está formado por los dos resultados posibles, cara y cruz. En el experimento aleatorio consistente en medir la altura de un sujeto, el espacio muestral está formado por todos los posibles resultados de la medición; si el experimento aleatorio consiste en lanzar una moneda dos veces, el espacio muestral está formado por cuatro posibles resultados: cara-cara, cara-cruz, cruz-cara, cruz-cruz; si se miden las alturas de dos sujetos, el espacio muestral está formado por todas las combinaciones resultantes de combinar las dos mediciones; etc.

Un **suceso** (S) es un subconjunto de un espacio muestral. Un **suceso simple** o **elemental** está formado por un único resultado (por ejemplo, obtener "cara-cara" en dos lanzamientos de una moneda). Un **suceso compuesto** está formado por más de un resultado (por ejemplo, obtener "una cara" en dos lanzamientos; es decir, obtener "cara-cruz" o "cruz-cara"). Al suceso formado por todos los resultados del espacio muestral se le llama **suceso seguro**; y a los resultados que no forman parte del espacio muestral, **suceso imposible**.

La **unión** ($\cup$) de dos sucesos es el conjunto de resultados distintos que forman parte de uno u otro suceso. La **diferencia** entre dos sucesos es el conjunto de resultados que pertenecen al primer suceso y no al segundo. La **intersección** ($\cap$) de dos sucesos es el conjunto de resultados que forman parte tanto de uno como de otro suceso. Dos sucesos se consideran **iguales** cuando incluyen los mismos resultados; y **exclusivos** cuando no tienen ningún resultado en común. Un suceso tiene su **complementario** en todos los resultados del espacio muestral que no forman parte de él.

Concepto de probabilidad

Existen diferentes formas de entender el concepto de probabilidad. Una aproximación intuitiva al concepto consiste en considerarlo como sinónimo de lo *fácil* o *difícil* que es observar cada uno de los sucesos de un espacio muestral. Si lanzamos al aire tres monedas, el suceso "tres caras" únicamente puede ocurrir de una manera: cara-cara-cara; sin embargo, el suceso "una cara" puede ocurrir de tres maneras distintas: cara-cruz-cruz, cruz-cara-cruz, cruz-cruz-cara. Por tanto, parece más *fácil* (más probable) observar el suceso "una cara" que el suceso "tres caras". Pero la probabilidad de un suceso es algo más que lo fácil o difícil que es observarlo: es un número que intenta cuantificar lo fácil o difícil que es observarlo.

El punto de vista **a priori**, también llamado *clásico*, asume que todos los sucesos elementales de un espacio muestral tienen las mismas posibilidades de ocurrir (principio de *indiferencia*) y cuantifica la probabilidad asociada a un suceso concreto (S) como su *frecuencia relativa teórica*:

$$P(S) = \frac{n_s}{n} \qquad [2.1]$$

es decir, como el número de resultados favorables al suceso (n_s) dividido entre el número de resultados posibles (n).

En este punto de vista se asume, por ejemplo, que los dos resultados posibles del lanzamiento de una moneda (cara y cruz) tienen las mismas posibilidades de ocurrir (es decir, son equiprobables); consecuentemente, la probabilidad *a priori* de cada uno de ellos vendrá dada por $P(\text{cara}) = P(\text{cruz}) = 1/2 = 0{,}5$. Del mismo modo, puesto que el suceso "cara-cara en dos lanzamientos" es uno entre cuatro posibles (cara-cara, cara-cruz, cruz-cara, cruz-cruz) que se asumen equiprobables, su probabilidad *a priori* vendrá dada por $P(\text{cara-cara}) = 1/4 = 0{,}25$.

El punto de vista **a posteriori**, también llamado *frecuentista* o *estadístico*, concibe la probabilidad de un suceso como el *límite al que tiende su frecuencia relativa*:

$$P(S) = \lim_{n \to \infty} \frac{n_s}{n} \qquad [2.2]$$

(aquí, n no es el número de sucesos del espacio muestral, sino el número de veces que se realiza el experimento aleatorio). Este punto de vista no hace ninguna suposición previa sobre las probabilidades de los sucesos; en lugar de eso, la probabilidad que se asigna a un suceso es su frecuencia relativa empírica, es decir, la proporción de veces que se observa el suceso al realizar el experimento aleatorio un número infinito de veces. Ahora bien, como no es posible realizar un experimento un número infinito de veces, la probabilidad *a posteriori* de un suceso hay que estimarla realizando el experimento *muchas veces*, tantas como sea necesario hasta observar que el valor de su frecuencia relativa se estabiliza. Imaginemos, por ejemplo, que se lanza una moneda 100 veces y que se obtienen 54 caras, es decir: $P(\text{cara}) = 54/100 = 0{,}54$; se sigue lanzando hasta 500 veces y se obtienen 242 caras: $P(\text{cara}) = 242/500 = 0{,}484$; se lanza 1.000 veces y se obtienen 511 caras: $P(\text{cara}) = 511/1.000 = 0{,}511$; se lanza 10.000 veces y se obtienen 4.962 caras: $P(\text{cara}) = 4.942/10.000 = 0{,}4962$; se lanza 20.000 veces y se obtienen 10.062 caras: $P(\text{cara}) = 10.062/20.000 = 0{,}5031$; se lanza 30.000 veces y se obtienen 14.967 caras: $P(\text{cara}) = 14.967/30.000 = 0{,}4989$; etc. Lo que ocurre al proceder de esta manera es que conforme va aumentando el número de ensayos (lanzamientos), la frecuencia

relativa del suceso *cara* se va estabilizando en torno a 0,50. Pues bien, ésta es la probabilidad *a posteriori* del suceso *cara*.

En la práctica, ambas formas de entender la probabilidad (*a priori* y *a posteriori*) son útiles y, también, necesarias. Por ejemplo, cuando se selecciona una muestra aleatoria de una población se está asumiendo que todos los elementos poblacionales tienen la misma probabilidad de ser elegidos (principio de indiferencia), es decir, se está adoptando un punto de vista *a priori*. Sin embargo, para conocer la probabilidad de que una persona de esa población sea hombre o fumador o tenga nivel de estudios superiores o una altura por encima de 180 cm, etc., no puede asumirse el principio de indiferencia (es decir, no puede asumirse que hay el mismo número de hombres que de mujeres o el mismo número de fumadores que de no fumadores, etc.); a no ser que se tenga información sobre todos los elementos de la población, las probabilidades asociadas a esos sucesos sólo pueden estimarse *a posteriori*, es decir, observando sus frecuencias relativas empíricas.

Sin embargo, adoptar uno u otro punto de vista no tiene implicaciones relevantes sobre las conclusiones a las que puede llegarse. Puesto que tanto las probabilidades *a priori* como las *a posteriori* se conciben como frecuencias relativas (teóricas en el primer caso y empíricas en el segundo), sus propiedades son idénticas:

1. La probabilidad de todos los sucesos del espacio muestral (el suceso seguro) vale 1. Es decir, $P(E) = 1$.
2. La probabilidad de un suceso es siempre no negativa. Es decir, $P(S) \geq 0$.
3. La probabilidad de la unión de dos o más sucesos mutuamente exclusivos es igual a la suma de las probabilidades individuales de cada suceso. Es decir, $P(S_1 \cup S_2 \cup S_3 \cup \cdots) = $ $= P(S_1) + P(S_2) + P(S_3) + \cdots$

Estas propiedades son precisamente las que han servido para formular una teoría **axiomática** o *formal* de la probabilidad. Adoptándolas como axiomas (pues son propiedades inherentes a cualquier probabilidad, ya sea ésta concebida *a priori* o *a posteriori*) y procediendo a partir de ellas por deducción se obtienen una serie de teoremas o reglas que constituyen lo que se conoce como *cálculo de probabilidades*. De estas reglas destacaremos dos particularmente útiles: la regla o teorema de la *multiplicación* (referida a la intersección de sucesos) y la regla o teorema de la *suma* (referida a la unión de sucesos).

Regla de la multiplicación

Entre los conceptos más interesantes que podemos encontrar en la teoría de la probabilidad se encuentra el de **probabilidad condicional**. Se refiere a la probabilidad de que ocurra un suceso cuando se impone la condición de que haya ocurrido otro previamente. Se representa mediante $P(S_1 \mid S_2)$ y se lee como "probabilidad condicional de S_1 dado S_2" o, simplemente, como "probabilidad de S_1 dado S_2".

Para entender fácilmente el significado de una probabilidad condicional, consideremos el ejemplo propuesto en la Tabla 2.1. Los resultados que muestra la tabla se han obtenido al clasificar a las 10.000 personas de una determinada población utilizando los criterios *sexo* (hombres, mujeres) y *tabaquismo* (fumadores, no fumadores).

De acuerdo con la ecuación [2.1] (número de casos favorables dividido entre el número de casos posibles), la probabilidad de que un sujeto elegido al azar sea *fumador*, es decir, la

probabilidad del suceso *fumador* (F), asumiendo que cualquier sujeto tiene la misma probabilidad de ser elegido, vale

$$P(F) = \frac{3.500}{10.000} = 0{,}35$$

Y la probabilidad de que un sujeto elegido al azar sea *hombre* (H) vale

$$P(H) = \frac{4.000}{10.000} = 0{,}40$$

Tabla 2.1. Frecuencias conjuntas de *sexo* y *tabaquismo*

	Fumadores	No fumadores	*Total*
Hombres	1.000	3.000	4.000
Mujeres	2.500	3.500	6.000
Total	3.500	6.500	10.000

Ahora bien, si se impone la condición de que el sujeto elegido sea *hombre*, entonces ¿cuál es la probabilidad de que sea *fumador*? Es decir, ¿cuál es la *probabilidad condicional* del suceso *fumador* dado el suceso *hombre*? Para responder a esta pregunta hay que tener en cuenta que los casos favorables, es decir, los *hombres fumadores*, son 1.000, y que, debido a la restricción impuesta, los casos posibles son 4.000 (sólo los hombres). Por tanto:

$$P(F|H) = \frac{1.000}{4.000} = 0{,}25$$

El numerador de esta probabilidad condicional recoge los 1.000 *hombres fumadores*, es decir, los elementos que forman parte de la intersección entre el suceso *fumador* y el suceso *hombre* ($F \cap H$). La probabilidad de esta combinación de sucesos (*ser fumador* **y** *ser hombre*) vale:

$$P(F \cap H) = \frac{1.000}{10.000} = 0{,}10$$

Y el denominador de la probabilidad condicional recoge los 4.000 elementos del suceso dado (H) cuya probabilidad ya sabemos que vale 0,40 (ver más arriba). En consecuencia:

$$P(F|H) = \frac{P(F \cap H)}{P(H)} = \frac{1.000/10.000}{4.000/10.000} = \frac{1.000}{4.000} = 0{,}25$$

Es decir, la *probabilidad condicional del suceso S_1 dado el suceso S_2 es igual a la probabilidad de la intersección de ambos sucesos dividida entre la probabilidad del suceso dado*:

$$P(S_1|S_2) = \frac{P(S_1 \cap S_2)}{P(S_2)} \qquad [2.3]$$

Precisamente esta definición de probabilidad condicional, que contiene en el numerador la probabilidad de la intersección de los dos sucesos, sirve para formular la **regla de la multiplicación** (también llamada regla del *producto*):

La probabilidad de la intersección de dos sucesos es igual a la probabilidad individual de uno de ellos multiplicada por la probabilidad condicional del otro.

Es decir,

$$P(S_1 \cap S_2) = P(S_2)\,P(S_1 | S_2) = P(S_1)\,P(S_2 | S_1) \quad [2.4]$$

Por tanto, hablar de *intersección* en el contexto de los sucesos de un espacio muestral es equivalente a hablar de *multiplicación* en el contexto de las probabilidades de esos sucesos.

Pero la definición [2.4] necesita ser matizada. Es claro que no todo suceso tiene por qué alterar la probabilidad de cualquier otro. De hecho, muchos sucesos no alteran las probabilidades de otros muchos. Pues bien, cuando dos sucesos no ven alteradas sus respectivas probabilidades individuales por la presencia del otro, decimos que esos sucesos son **independientes**. Cuando se da esta circunstancia, la probabilidad condicional de un suceso no difiere de su probabilidad individual. Es decir, si dos sucesos son independientes se verifica

$$P(S_1 | S_2) = P(S_1) \quad [2.5]$$

Por tanto, si dos sucesos son *independientes*, la regla de la multiplicación ya presentada más arriba, se simplifica:

La probabilidad de la intersección de dos sucesos independientes es igual al producto de sus probabilidades individuales. Y a la inversa: si la probabilidad de la intersección de dos sucesos es igual al producto de sus probabilidades individuales, entonces esos sucesos son independientes.

Volviendo a los datos de la Tabla 2.1, ¿puede decirse que el suceso *hombre* es independiente del suceso *fumador*? Sabemos (ver más arriba) que la probabilidad de la intersección entre esos sucesos vale 0,10, la del suceso *hombre* 0,40 y la del suceso *fumador* 0,35. Si los dos sucesos fueran independientes, la probabilidad de su intersección (0,10) debería ser igual al producto de sus probabilidades individuales ($0{,}40 \times 0{,}35 = 0{,}14$). Puesto que la probabilidad 0,10 es distinta de la probabilidad 0,14, podemos decir que los sucesos *hombre* y *fumador* no son independientes.

Regla de la suma

Si dos sucesos son mutuamente exclusivos (es decir, si no tienen elementos en común; ver Figura 2.1, gráfico de la izquierda), la probabilidad de su unión es la suma de sus probabilidades individuales. Esto es lo que afirma el axioma 3. Ahora bien, si los sucesos no son exclusivos (es decir, si tienen algún elemento en común; ver Figura 2.1, gráfico de la derecha), a la probabilidad de la unión hay que restarle la parte que tienen en común, es decir, la intersección de ambos. Este razonamiento da pie para formular la **regla de la suma**:

Si S_1 y S_2 son sucesos *exclusivos*: $P(S_1 \cup S_2) = P(S_1) + P(S_2)$.
Si S_1 y S_2 son sucesos *no exclusivos*: $P(S_1 \cup S_2) = P(S_1) + P(S_2) - P(S_1 \cap S_2)$. [2.6]

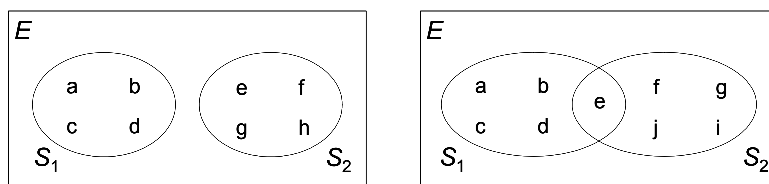
Por tanto, hablar de *unión* en el contexto de los sucesos de un espacio muestral es equivalente a hablar de *suma* en el contexto de las probabilidades de esos sucesos. En el ejemplo de la

Figura 2.1, la probabilidad de la unión de los dos sucesos del gráfico de la izquierda se obtiene sumando las probabilidades individuales de ambos sucesos. Sin embargo, en el gráfico de la derecha, la probabilidad de la unión de ambos sucesos no se corresponde con la suma de las probabilidades individuales; a la suma de esas probabilidades individuales hay que restar la probabilidad de la intersección (es decir, hay que restar la probabilidad correspondiente al elemento *e*, pues se ha sumado dos veces).

Volviendo a los datos de la Tabla 2.1, la regla de la suma puede utilizarse para conocer la probabilidad de la unión de los sucesos *hombre* y *fumador*, es decir, la probabilidad de que un sujeto elegido al azar sea *hombre* **o** *fumador*. Obviamente, ser *hombre* y ser *fumador* no son sucesos exclusivos, pues una persona puede ser al mismo tiempo ambas cosas. Por tanto:

$$P(H \cup F) = P(H) + P(F) - P(H \cap F) = 0{,}40 + 0{,}35 - 0{,}10 = 0{,}65$$

Figura 2.1. Sucesos *exclusivos* (izquierda) y *no exclusivos* (derecha) en el espacio muestral *E*



Combinando la regla de la multiplicación y la regla de la suma se llega a un teorema, muy conocido en estadística, llamado **teorema de Bayes**. No obstante, puesto que no ayuda a resolver nada que no se resuelva con las dos reglas estudiadas, no será tratado aquí (el lector interesado puede consultar, por ejemplo, Amón, 1984, págs. 53-59).

Apéndice 2

Combinatoria (reglas de contar)

Utilizar el cálculo de probabilidades requiere, entre otras cosas, conocer el espacio muestral con el que se desea trabajar, es decir, los posibles resultados del correspondiente experimento aleatorio. Aunque con espacios muestrales pequeños es fácil calcular el número total de resultados, con espacios muestrales grandes la tarea se complica bastante. En estos casos es muy útil disponer de alguna herramienta que facilite el trabajo. Como también lo es contar con herramientas que ayuden a calcular, por ejemplo, cuántas comparaciones por pares pueden hacerse con un determinado número de elementos; o de cuántas maneras distintas puede ordenarse un conjunto de estímulos para presentarlos a una muestra de sujetos. Todos estos cálculos pueden realizarse fácilmente con las llamadas *reglas de contar*, algunas de las cuales se describen en este apartado.

Comencemos con el **principio fundamental de la combinatoria**. Sirve para resolver muchas de las situaciones que podemos encontrarnos y es muy fácil de aplicar:

> Si el suceso S_1 puede ocurrir de n_1 maneras, el suceso S_2 de n_2 maneras, ..., el suceso S_k de n_k maneras, los k sucesos S_1, S_2, ..., S_k pueden ocurrir conjuntamente de $n_1 \times n_2 \times \cdots \times n_k$ maneras.

Para comprender la utilidad de este principio, vamos a comenzar con un par de ejemplos de juegos que pueden resultar bastante familiares. Primero: ¿cuántos resultados posibles tiene una quiniela de fútbol? Una quiniela tiene 15 resultados, cada uno de los cuales puede ocurrir de 3 maneras distintas; por tanto, los 15 resultados juntos pueden ocurrir de $3 \times 3 \times \cdots \times 3 = 3^{15} = 14.348.907$ maneras distintas. Segundo: ¿cuántos resultados hay en la lotería primitiva? En este juego se eligen al azar 6 números entre 49 posibles (números del 1 al 49); el primer número elegido puede ser uno cualquiera de los 49 posibles; el segundo, uno de 48 restantes (pues el segundo resultado no puede ser el número que ya ha salido como primer resultado); el tercero, uno de 47 restantes; ...; el sexto, uno de los 44 restantes; por tanto, los 6 números elegidos pueden aparecer de $49 \times 48 \times 47 \times 46 \times 45 \times 44 = 10.068.347.520$ maneras.

Aunque ambos casos se resuelven utilizando la misma estrategia, lo cierto es que difieren en un aspecto importante. En el caso de la quiniela, cada posible resultado es distinto de cada otro porque el orden en el que aparecen las quince apuestas es crucial. En la lotería primitiva, sin embargo, no todos los posibles resultados son distintos entre sí, sino que hay algunos que son equivalentes a otros; por ejemplo, el resultado $\{1, 2, 3, 4, 5, 6\}$ es, obviamente, equivalente al resultado $\{1, 3, 5, 2, 4, 6\}$; y también es equivalente a cualquier otro que contenga los mismos números aunque estén en distinto orden. Por tanto, para calcular correctamente los posibles resultados de la lotería primitiva es necesario tener en cuenta de cuántas maneras pueden ordenarse 6 números distintos. Veamos: el primer número puede ocupar cualquiera de las 6 posiciones disponibles; el segundo, cualquiera de las cinco restantes; ...; el sexto, la única disponible al final. Aplicando el principio fundamental de la combinatoria se llega a la conclusión de que 6 números distintos pueden ordenarse de $6 \times 5 \times 4 \times 3 \times 2 \times 1 = 720$ maneras distintas. Dado que esto ocurre con cualquier combinación de 6 números, el número de posibles resultados *distintos* en la lotería primitiva vendrá dado por el cociente entre el primer cálculo realizado (10.068.347.520, cantidad que incluye muchos resultados equivalentes) y las distintas maneras de ordenar 6 números (720), es decir, 13.983.816 maneras distintas.

Para terminar de aclarar estas diferencias entre posibles resultados de un experimento aleatorio, consideremos un ejemplo algo más simple. Supongamos que lanzamos una moneda dos veces y observamos el resultado. Llamando *c* al resultado *cara* y *x* al resultado cruz, este experimento aleatorio tiene asociados cuatro posibles resultados: *cc*, *cx*, *xc*, *xx*. El hecho de que estos cuatro resultados se consideren o no *distintos entre sí* dependerá del criterio que se aplique para distinguirlos: (1) si se considera que dos resultados son distintos tanto si contienen elementos distintos como si, conteniendo los mismos, se encuentran en distinto orden, entonces los 4 resultados son distintos; (2) si se considera que dos resultados son distintos únicamente si contienen elementos distintos, entonces hay 3 resultados distintos: *cc*, *cx*, *xx* (los resultados *cx* y *xc* cuentan como un único resultado); (3) por último, si se considera que dos resultados son distintos únicamente cuando contienen los mismos elementos pero en distinto orden, entonces hay 2 resultados distintos: *cx* y *xc*. A los resultados de aplicar el primer criterio se les llama *variaciones*; a los de aplicar el segundo criterio, *combinaciones*; y a los de aplicar el tercer criterio, *permutaciones*. Y, aunque todos estos resultados pueden calcularse utilizando el principio fundamental de la combinatoria, existen algunas fórmulas que facilitan el trabajo[7].

Variaciones ($V_{N,n}$): número de grupos distintos que es posible formar con N elementos tomados de n en n, considerando que dos grupos son distintos tanto si difieren en alguno de sus elementos como si difieren en el orden de los mismos[8]:

$$V_{N,n} = \frac{N!}{(N-n)!} \qquad [2.7]$$

[7] También pueden formarse variaciones, combinaciones y permutaciones con repetición, pero su utilidad para el analista de datos es más bien escasa y no serán tratadas aquí. El lector interesado en ellas puede consultar Amón (1979, pág. 33).

[8] El signo "!" se lee *factorial* ($n!$ se lee *n factorial*; 5! se lee *cinco factorial*) y significa que el número que le precede hay que multiplicarlo por todos los números enteros menores que él hasta llegar a 1. Así, $5! = 5 \times 4 \times 3 \times 2 \times 1 = 120$. La excepción a esta regla la constituye el número 0: se asume que $0! = 1$.

Supongamos que 10 candidatos optan a 3 puestos de trabajo con diferente remuneración. ¿De cuántas maneras distintas pueden repartirse los 3 puestos entre los 10 candidatos? Para responder a esta pregunta es necesario tener en cuenta que cada grupo de tres candidatos es distinto de cada otro tanto si incluye algún candidato distinto como si los puestos se reparten de forma distinta entre los mismos tres candidatos (importa el orden). Por tanto, se trata de variaciones de 10 elementos tomados de 3 en 3:

$$V_{10,\,3} \;=\; \frac{10!}{(10-3)!} \;=\; 10\times 9\times 8 \;=\; 720 \;\; \text{maneras}$$

Utilizando el principio fundamental de la combinatoria se llega al mismo resultado: el primer premio puede recaer sobre 10 personas distintas, el segundo sobre 9 y el tercero sobre 8; por tanto, los tres premios pueden repartirse de $10 \times 9 \times 8 = 720$ maneras distintas.

Combinaciones ($C_{N,\,n}$): número de grupos distintos que es posible formar con N elementos tomados de n en n, considerando que dos grupos son distintos únicamente si difieren en alguno de sus elementos:

$$C_{N,\,n} = \binom{N}{n} = \frac{N!}{n!\,(N-n)!} \qquad [2.8]$$

Supongamos que tenemos que formar grupos de trabajo de 3 personas con los 10 empleados de un departamento. ¿Cuántos grupos distintos de 3 personas pueden formarse? Obviamente, dos grupos serán distintos únicamente si no contienen las mismas personas; aquí, el orden en el que se elije a las personas no afecta a la composición del grupo. Por tanto, se trata de combinaciones de 10 elementos tomados de 3 en 3:

$$C_{10,\,3} = \binom{10}{3} = \frac{10!}{3!\,(10-3)!} = \frac{10\times 9\times 8}{3\times 2\times 1} = 120 \;\; \text{grupos}$$

Utilizando el principio fundamental de la combinatoria se llega al mismo resultado: el primer miembro del grupo puede ser uno cualquiera de los 10 empleados; el segundo, uno cualquiera de los 9 restantes; el tercero, uno cualquiera de los 8 restantes. Por tanto, con los 10 empleados es posible formar un total de $10 \times 9 \times 8 = 720$ grupos. Pero, como muchos de estos grupos son equivalentes (están formados por los mismos sujetos aunque en distinto orden), la cantidad obtenida (720) hay que dividirla entre el número de ordenaciones distintas que es posible hacer con tres elementos: $3 \times 2 \times 1 = 6$. En consecuencia, es posible formar un total de $720/6 = 120$ grupos distintos.

Permutaciones (P_n): número de ordenaciones distintas que es posible realizar con n elementos:

$$P_n \;=\; n! \qquad [2.9]$$

Por ejemplo, ¿de cuántas maneras distintas pueden asignarse los 10 empleados del ejemplo anterior a los 10 despachos disponibles en el departamento? La solución, ahora, no consiste en hacer subgrupos, sino en ordenar a los 10 empleados de todas las formas posibles. Se trata, por tanto, de permutaciones de 10 elementos:

$$P_{10} \;=\; 10\,! = 10\times 9\times 8\times \cdots \times 1 \;=\; 3.628.800 \;\; \text{maneras distintas}$$

Utilizando el principio fundamental de la combinatoria se obtiene el mismo resultado: el primer miembro del grupo puede ocupar uno cualquiera de los 10 despachos disponibles; el segundo, uno cualquiera de los 9 restantes; ...; el décimo, el único despacho disponible; por tanto, los 10 empleados pueden repartirse en los 10 despachos de $10 \times 9 \times 8 \times \cdots \times 1 = 3.628.800$ maneras distintas.

En lo que a nosotros más nos interesa, tanto las variaciones como las combinaciones tienen la importante utilidad de permitir calcular el número de muestras distintas que es posible extraer de una población finita. Supongamos que se extrae una muestra de $n = 5$ personas de una población de $N = 20$ personas (si la población tuviera 20 millones de personas el razonamiento sería el mismo). Ciertamente, un grupo de personas no cambia porque las mismas 5 personas se elijan en un orden u otro. Pero, cuando se extraen muestras aleatorias, lo que interesa es que cualquiera de ellas tenga la misma probabilidad de ser elegida. Y puesto que los elementos pueden aparecer en distinto orden, cada una de esas posibilidades tendrá asociada una probabilidad. Por tanto, desde este punto de vista, una muestra debe considerarse distinta de otra tanto si contiene algún elemento distinto como si, conteniendo los mismos, se encuentran en distinto orden. Consiguientemente, el número de muestras posibles vendrá dado por las variaciones de 20 elementos tomados de 5 en 5:

$$V_{20,5} = \frac{20!}{(20-5)!} = 20 \times 19 \times 18 \times 17 \times 16 = 1.860.480 \text{ muestras posibles}$$

Ahora bien, si se considera que una muestra es distinta de otra únicamente cuando contiene algún elemento distinto, entonces el número de muestras posibles vendrá dado por las combinaciones de 20 elementos tomados de 5 en 5:

$$C_{20,5} = \frac{20!}{5!(20-5)!} = \frac{20 \times 19 \times 18 \times 17 \times 16}{5 \times 4 \times 3 \times 2 \times 1} = \frac{1.860.480}{120} = 15.504 \text{ muestras distintas}$$

Es claro que el número de muestras posibles que resulta con uno y otro criterio es muy distinto. Sin embargo, la probabilidad asociada a cada posible muestra es la misma independientemente del criterio adoptado. En el primer caso (variaciones), esa probabilidad vale uno dividido entre las 1.860.480 muestras posibles; en el segundo, uno dividido entre las 15.504 muestras posibles.

Cómo seleccionar una muestra aleatoria

Al trabajar con poblaciones finitas, la extracción de una muestra aleatoria requiere, en general, como primer paso, que los elementos poblacionales estén identificados de alguna manera. Una forma apropiada de identificarlos consiste en numerar los elementos poblacionales de 1 a N y, a continuación, utilizar una *tabla de números aleatorios* para elegir los elementos que formarán parte de la muestra.

Las tablas de números aleatorios (como la tabla A del apéndice final) han sido elaboradas de tal forma que todos los dígitos del 0 al 9 aparecen con la misma frecuencia y repartidos de forma aleatoria (los dígitos suelen aparecer en estas tablas formando grupos para facilitar su lectura, pero esa agrupación no tiene otro significado).

Para ilustrar cómo utilizar la tabla de números aleatorios, supongamos que tenemos que extraer una muestra de tamaño $n = 50$ de una población de tamaño $N = 800$. El primer paso consiste en numerar los elementos poblacionales de 1 a 800 (normalmente se trabaja con listas que tienen resuelto esto). A continuación, en la tabla de números aleatorios (la del apéndice final está formada por 1.000 dígitos: 40 filas por 25 columnas) seleccionamos al azar un dígito cualquiera. Supongamos que la elección recae sobre el dígito colocado en la 29ª fila y en la 13ª columna: hemos elegido el número 5. Leyendo a partir de esa posición de izquierda a derecha (aunque podría hacerse en cualquier otra dirección) encontramos los siguientes números de tres dígitos (tres dígitos porque ése es el número de dígitos del tamaño poblacional: 800): 541, 149, 050, etc. Seguimos así hasta obtener los 50 elementos que deben formar parte de la muestra. Si reanudamos la secuencia donde la hemos dejado, el siguiente número es 944; como este número es mayor que 800 (tamaño poblacional), desechamos ese valor y continuamos: 109, 341, etc. Por supuesto, se puede continuar indistintamente en la fila de abajo o en la de arriba; cualquier dirección que se tome ofrecerá una secuencia aleatoria.

El problema de los métodos de extracción basados en tablas de números aleatorios es que sólo resultan aplicables cuando se está trabajando con poblaciones finitas. En una población infinita no es posible, por ejemplo, numerar todos los elementos que la componen. En estos casos es necesario adoptar una estrategia de muestreo diferente. Una de estas estrategias se conoce con el nombre de *simulación*: "técnica de muestreo estadístico controlado utilizada, junto con un modelo, para obtener respuestas aproximadas sobre problemas probabilísticos (...) complejos" (Lewis y Orav, 1989, pág. 9). En el apéndice del Capítulo 6 se ofrece una breve explicación de un método de simulación conocido como método Monte Carlo.

Ejercicios

2.1. En el ejercicio 1.1 (ver capítulo anterior) hemos propuesto un conjunto de características con el objetivo de aprender a identificar el nivel de medida que era posible alcanzar con ellas. Ahora se trata de decidir si esas características, a las que ya podemos empezar a llamar variables, deben ser clasificadas como *categóricas* o como *cuantitativas*.

a. Percepción subjetiva del dolor.
b. Grupo de tratamiento (experimental, control).
c. Satisfacción con un determinado servicio.
d. Peso de los recién nacidos.
e. Tiempo de reacción.
f. Calidad percibida del estado de salud general.
g. Rendimiento en el test de inteligencia Raven.
h. Actitud hacia el aborto (en contra, indiferente, a favor).
i. Rendimiento en una prueba de cálculo numérico.
j. Nivel socioeconómico (bajo, medio, alto).
k. Número de aciertos en una prueba de rendimiento.
l. Calidad del material recordado.
m. Nivel de ansiedad.
n. Intensidad del ruido ambiental.
ñ. Años de experiencia educativa de un profesor.
o. Color de un estímulo (rojo, amarillo, verde, azul).
p. Dosis de un fármaco (0 mg, 100 mg, 250 mg, 500 mg).
q. Grado de dificultad de una pregunta.
r. Nivel de alcohol en sangre (g/l).
s. Consumo de alcohol (nulo, bajo, medio, alto).
t. Número de cigarrillos/día.
u. Tabaquismo (fumadores, exfumadores, no fumadores).
v. Puntuaciones en la escala de depresión de Hamilton.
w. Número de accidentes de tráfico ocurridos en fin de semana.
x. Tipo de ideología política (izquierda, centro, derecha).
y. Nivel de conservadurismo medido en el continuo izquierda-derecha.
z. Tipo de tratamiento antidepresivo (farmacológico, psicológico, mixto).

2.2. A continuación se ofrecen varias afirmaciones que pueden ayudar a precisar el significado de algunos de los conceptos introducidos en este capítulo. ¿Cuál de ellas es verdadera y cuál falsa?

a. Un parámetro es una característica individual de cada elemento de una población.
b. Un estadístico es un número y, por tanto, una constante.
c. Al seleccionar varias muestras de una misma población y calcular en cada una de ellas un estadístico, el valor de ese estadístico será siempre el mismo sólo si las muestras son aleatorias y del mismo tamaño.
d. Bajo ciertas circunstancias, los estudiantes de la Universidad Autónoma de Madrid constituyen una población.
e. Una muestra aleatoria de los estudiantes de un colegio de una ciudad puede ser considerada representativa de los estudiantes de esa ciudad.

2.3. En un ensayo clínico diseñado para probar la eficacia de un nuevo fármaco destinado a pacientes con insomnio se utiliza una muestra de los pacientes con insomnio que acuden a la consulta de un determinado hospital durante un determinado periodo de tiempo. Señalar la(s) alternativa(s) correcta(s):

a. Se tiene una muestra aleatoria de pacientes con insomnio.
b. Se tiene una muestra no aleatoria de pacientes con insomnio.
c. La población de referencia es la de pacientes con insomnio.

2.4. Para estudiar la relación entre las variables tabaquismo y enfisema pulmonar se han recogido datos en tres hospitales de la zona sur de Madrid. Al comienzo del estudio, los sujetos, elegidos aleatoriamente entre los pacientes sin enfisema que han acudido a consulta durante un año, se han clasificado como fumadores, exfumadores y no fumadores. Tras diez años de seguimiento se ha registrado la presencia o no de enfisema pulmonar.

a. ¿Cuál es la población de referencia?
b. ¿Cuál es el parámetro que interesa estudiar?
c. ¿Se ha seleccionado una muestra aleatoria de la población de referencia?
d. ¿A qué tipo de conclusión permite llegar un estudio de estas características (descriptiva, relacional, explicativa)?

2.5. Señalar cuáles de las siguientes afirmaciones son verdaderas y cuáles son falsas:

a. Si dos sucesos son independientes, la probabilidad de uno de ellos es la misma tanto si el otro suceso está presente como si no.
b. Si dos sucesos son independientes, la probabilidad de su suma es igual a la suma de sus probabilidades.
c. Si dos sucesos son exclusivos, su probabilidad conjunta es igual al producto de sus probabilidades individuales.
d. Si se lanza una moneda al aire cinco veces y en las cinco ocasiones sale cara, la probabilidad de que salga cara en el sexto lanzamiento es menor que la probabilidad de que salga cruz (asumimos que la moneda no está trucada y que el lanzamiento es imparcial).
e. Si se lanza una moneda al aire diez veces, el resultado "5 caras" es igual de probable que el resultado "7 caras".

2.6. Un examen consta de tres preguntas. Todas ellas tienen cinco alternativas de respuesta de las que sólo una es correcta. Si un sujeto responde al azar, ¿cuál es la probabilidad de que:

a. No acierte ninguna pregunta?
b. Acierte una pregunta?
c. Acierte dos preguntas?
d. Acierte las tres preguntas?

2.7. En un estudio sobre discriminación visual se presentan a un sujeto 10 pares de estímulos luminosos de la misma intensidad. La tarea consiste en decidir si los estímulos de cada par tienen o no la misma intensidad. Si el sujeto realiza la tarea respondiendo al azar:

a. ¿Cuál es la probabilidad de que no dé la respuesta correcta en ningún par?
b. ¿Cuál es la probabilidad de que dé la respuesta correcta en un solo par?

2.8. En un estudio dirigido a establecer la prevalencia de la demencia senil en personas mayores de 65 años, se han recogido datos sobre 5.000 personas. Cada persona se ha clasificado utilizando dos criterios: *sexo* (hombre, mujer) y *demencia senil* (sí, no). La siguiente tabla muestra los resultados obtenidos:

	Demencia senil		
Sexo	*S* = sí	*N* = no	*Total*
H = hombres	500	1.500	2.000
M = mujeres	750	2.250	3.000
Total	1.250	3.750	5.000

a. ¿Son independientes los sucesos *ser hombre* y *padecer demencia*?
b. Si se elige una persona al azar, ¿cuál es la probabilidad de que se trate de una mujer que no padece demencia?
c. Si se elige una persona al azar y resulta ser hombre, ¿cuál es la probabilidad de que padezca demencia?
d. Si se elige al azar una persona y resulta que padece demencia, ¿cuál es la probabilidad de que sea un hombre?

2.9. Supongamos que la población de personas mayores de 60 años está formada por un 40% de hombres (*H*) y un 60% de mujeres (*M*). Supongamos, además, que el porcentaje de personas dependientes (*D*) en esa población es del 10% entre los hombres y del 20% entre las mujeres. Si se elige una persona al azar:

a. ¿Cuál es la probabilidad de que la persona elegida sea un hombre dependiente?
b. ¿Cuál es la probabilidad de que la persona elegida sea una mujer dependiente?
c. ¿Cuál es la probabilidad de que la persona elegida sea dependiente?
d. Si la persona elegida es dependiente, ¿cuál es la probabilidad de que sea un hombre?

2.10. Tres pruebas diagnósticas para la detección del Alzheimer (*A*, *B* y *C*) detectan la enfermedad en el 90, 80 y 70%, respectivamente, de las personas que la padecen. Si el diagnóstico de cada prueba es independiente del de las demás:

a. ¿Cuál es la probabilidad de detectar la enfermedad si se aplican las pruebas A y B?
b. ¿Cuál es la probabilidad de no detectar la enfermedad si se aplican las pruebas B y C?
c. Si se considera que la enfermedad está presente sólo si las tres pruebas la detectan, ¿cuál es la probabilidad de que un enfermo de Alzheimer sea diagnosticado como tal?
d. Si se considera que la enfermedad está presente sólo si, aplicadas las tres pruebas, al menos dos de ellas la detectan, ¿cuál es la probabilidad de que un enfermo de Alzheimer sea diagnosticado como tal?

2.11. Consideremos dos preguntas de un examen: P_1 y P_2. Ambas tienen varias alternativas de respuesta de las que sólo una es correcta, pero la primera pregunta tiene cuatro alternativas y la segunda cinco. Un estudiante responde al azar a una de esas dos preguntas y acierta (A).

a. ¿Cuál es la probabilidad de que la pregunta respondida sea la primera?
b. ¿Cuál es la probabilidad de que la pregunta respondida sea la segunda?

2.12. Un detector de mentiras diagnostica correctamente al 90% de las personas que mienten (M) y al 95% de las que no mienten. Se elige al azar una persona de un colectivo de 100 personas del que se sabe que 20 mienten.

a. Tanto si esa persona miente como si no, ¿cuál es la probabilidad de que el detector ofrezca un diagnóstico correcto?
b. Si el detector indica que esa persona miente, ¿cuál es la probabilidad de que el diagnóstico sea correcto?

2.13. Se sabe que, en una determinada población, la prevalencia de una enfermedad concreta es del 30%. Se dispone de una prueba diagnóstica con una sensibilidad (= diagnóstico positivo cuando la persona padece la enfermedad) del 90% y una especificidad (= diagnóstico negativo cuando la persona no padece la enfermedad) del 80%. Al realizar un diagnóstico concreto a un sujeto de esa población:

a. ¿Cuál es la probabilidad de que la prueba dé un resultado positivo?
b. ¿Cuál es la probabilidad de que la prueba dé un diagnóstico equivocado?
c. Si la prueba da un resultado positivo, ¿cuál es la probabilidad de que la persona no esté enferma?

2.14. El 40% de los aspirantes a un puesto de trabajo ha superado (S) una determinada prueba de selección. El 80% de los aspirantes que superan esa prueba terminan siendo contratados (C), frente a sólo el 5% de los que no la superan. Si un aspirante es finalmente contratado, ¿cuál es la probabilidad de que haya superado la prueba de selección?

2.15. En un determinado instituto de enseñanza secundaria, el 48% de los estudiantes son chicos (O) y el 52% chicas (A). El 15% de los chicos tiene 19 años o más; el 5% de las chicas tiene 19 años o más.

a. ¿Cuál es la probabilidad de que un estudiante elegido al azar tenga 19 años o más?
b. ¿Son independientes los sucesos *ser chico* y *tener 19 años o más*?
c. Si se elige un estudiante al azar y resulta tener más de 19 años, ¿cuál es la probabilidad de que sea una chica?

2.16. El azar (la selección aleatoria) desempeña un importante rol en el análisis de datos: no sólo es la única estrategia de que disponemos para obtener muestras representativas, sino que las reglas que se derivan de la teoría de la probabilidad se basan en él. Pero el azar, o las leyes del azar, tienen otras muchas aplicaciones. Una de ellas, muy interesante, tiene que ver con garantizar el anonimato en los cuestionarios que contienen preguntas sensibles.

Imaginemos un estudio en el que se trata de obtener una estimación de la proporción de personas que defraudan a Hacienda (lo mismo valdría para conductas como el consumo de drogas, para los gustos o inclinaciones sexuales, etc.; es decir, para conductas que las personas se sienten inclinadas a ocultar o *maquillar* por ser socialmente indeseables). Lógicamente, pocas personas decidirán confesar un delito (como defraudar a Hacienda) a no ser que reciban plenas garantías de que su respuesta permanecerá en el anonimato.

Wonnacott y Wonnacott (1990, págs. 107-108) han propuesto una ocurrente manera de garantizar el anonimato de las respuestas a un cuestionario. La estrategia consiste en pedir al entrevistado que lance una moneda en privado con la siguiente indicación: (1) "si sale cara, responda a la pregunta: ¿ha defraudado alguna vez a Hacienda?"; (2) "si sale cruz, vuelva a lanzar la moneda y diga si ha vuelto a salir cruz". Al proceder de esta manera, si el entrevistado responde "sí", el entrevistador no tiene forma de saber si el entrevistado ha defraudado a Hacienda o ha obtenido dos cruces. El anonimato está garantizado.

Por supuesto, esta estrategia no permite conocer las respuestas individuales de los entrevistados. Pero sí permite obtener una estimación de la proporción de personas que defraudan a Hacienda.

a. Supongamos que la verdadera proporción (π) de entrevistados que defrauda a Hacienda vale 0,40. ¿Qué proporción (P) de respuestas "sí" cabe esperar encontrar en una muestra concreta con el procedimiento descrito?

b. Si en una muestra concreta se obtiene $P = 0{,}15$, ¿qué valor habrá que estimar para π?

c. Para poder responder a las dos preguntas anteriores es necesario asumir que ciertas cosas ocurren de cierta manera. ¿Qué cosas y de qué manera?

Soluciones

2.1. En estas soluciones se indica no sólo si una variable es categórica o cuantitativa. En las variables categóricas se indica si son nominales u ordinales. En las cuantitativas se distingue entre las que son típicamente cuantitativas (de intervalos o de razón; aclarando si son continuas o discretas) y las que, no siendo típicamente cuantitativas, tampoco son estrictamente ordinales.

a. Cuantitativa (no estrictamente ordinal).
b. Categórica (nominal).
c. Cuantitativa (no estrictamente ordinal).
d. Cuantitativa (continua).
e. Cuantitativa (continua).
f. Cuantitativa (no estrictamente ordinal).
g. Cuantitativa (no estrictamente ordinal).
h. Categórica (ordinal).
i. Cuantitativa (no estrictamente ordinal).
j. Categórica (ordinal).

k. Cuantitativa (discreta).
l. Cuantitativa (no estrictamente ordinal).
m. Cuantitativa (no estrictamente ordinal).
n. Cuantitativa (continua).
ñ. Cuantitativa (discreta).
o. Categórica (nominal).
p. Cuantitativa (discreta).
q. Cuantitativa (no estrictamente ordinal).
r. Cuantitativa (continua).
s. Categórica (ordinal).
t. Cuantitativa (discreta).
u. Categórica (nominal).
v. Cuantitativa (no estrictamente ordinal).
w. Cuantitativa (discreta).
x. Categórica (nominal).
y. Cuantitativa (no estrictamente ordinal).
z. Categórica (nominal).

2.2. La alternativa *a* es falsa: un parámetro es una característica de la población, no de cada uno de sus elementos. La alternativa *b* es falsa: un estadístico es una variable (su valor depende de la muestra concreta en la se calcula). La alternativa *c* es falsa: el valor de un estadístico varía de muestra a muestra (por eso decimos que es una variable). La alternativa *d* es verdadera: los estudiantes de la Universidad Autónoma constituyen una población cuando las conclusiones del estudio se refieren sólo a ellos. La alternativa *e* es falsa: sólo sería representativa de los estudiantes de ese colegio, que son los únicos que han tenido la posibilidad de ser elegidos.

2.3. Las tres alternativas son falsas. La muestra no es aleatoria porque los pacientes se van seleccionando conforme van llegando a la consulta; esto podría confundirse con una muestra aleatoria (argumentando que los pacientes llegan a la consulta aleatoriamente), pero no lo es porque no todos los pacientes con insomnio han tenido la posibilidad de ser elegidos; sólo los que han acudido a consulta durante el periodo de reclutamiento (esto descarta la alternativa *a*). La muestra, por tanto, no es aleatoria y no ha sido de la población de pacientes con insomnio, sino de la población de potenciales pacientes del hospital seleccionado.

2.4. *a*. La de potenciales pacientes de la zona de influencia de los hospitales seleccionados.
b. El grado de relación entre tabaquismo y enfisema pulmonar.
c. No. Es una muestra aleatoria de los pacientes que acuden a consulta.
d. Relacional. No existe asignación aleatoria a las condiciones del estudio y, por tanto, no hay forma de excluir el posible efecto de terceras variables.

2.5. *a*. Verdadera (ver ecuación [2.5]).
b. Falsa. Esto sólo será así si los sucesos son exclusivos.
c. Falsa. Esto sólo será así si los sucesos son independientes.
d. Falso. El resultado de cada lanzamiento es independiente del resultado de cualquier otro lanzamiento.
e. Falso. El resultado "5 caras" puede ocurrir de 252 maneras. El resultado "7 caras" puede ocurrir de 120 maneras (combinaciones sin repetición; ver Apéndice 2).

2.6. Puesto que las preguntas tienen cinco alternativas con sólo una correcta, la probabilidad de acertar una pregunta por azar vale: $P(A) = 1/5 = 0{,}20$; y la de no acertar: $P(nA) = 1 - 0{,}20 = 0{,}80$. Por tanto:

a. $0{,}80 \times 0{,}80 \times 0{,}80 = 0{,}512$.

b. Hay tres formas distintas de obtener un acierto: se puede acertar la primera pregunta, la segunda o la tercera. Por tanto, $0{,}20 \times 0{,}80 \times 0{,}80 + 0{,}80 \times 0{,}20 \times 0{,}80 + 0{,}80 \times 0{,}80 \times 0{,}20 = 0{,}384$.

c. Hay tres formas distintas de obtener dos aciertos: se puede fallar la primera pregunta, o la segunda, o la tercera. Por tanto, $0{,}80 \times 0{,}20 \times 0{,}20 + 0{,}20 \times 0{,}80 \times 0{,}20 + 0{,}20 \times 0{,}20 \times 0{,}80 = 0{,}096$.

d. $0{,}20 \times 0{,}20 \times 0{,}20 = 0{,}008$.

2.7. *a*. $0{,}50^{10} = 0{,}00098$.

b. Hay 10 formas distintas de obtener un acierto: se puede acertar el primer par, el segundo, el tercero, ..., o el décimo. Por tanto, $10 \times 0{,}50^{10} = 0{,}0098$.

2.8. *a*. Sí son independientes: $P(H \cap S) = P(H)P(S) = 500/5.000 = (2.000/5.000)(1.250/5.000) = 0{,}10$.

b. $P(M \cap N) = 2.250 / 5.000 = 0{,}45$.

c. $P(S \mid H) = P(S \cap H) / P(H) = 0{,}10 / (2.000/5.000) = 0{,}25$.

d. $P(H \mid S) = P(H \cap S) / P(S) = 0{,}10 / (1.250/5.000) = 0{,}40$.

2.9. Sabemos que $P(H) = 0{,}40$; $P(M) = 0{,}60$; $P(D|H) = 0{,}10$ y $P(D|M) = 0{,}20$. Por tanto:

a. $P(H \cap D) = P(H)\, P(D|H) = 0{,}40 \times 0{,}10 = 0{,}04$.

b. $P(M \cap D) = P(M)\, P(D|M) = 0{,}60 \times 0{,}20 = 0{,}12$.

c. $P(D) = P(H \cap D) + P(M \cap D) = 0{,}04 + 0{,}12 = 0{,}16$.

d. $P(H|D) = P(H \cap D) / P(D) = 0{,}04 / 0{,}16 = 0{,}25$.

2.10. *a*. La enfermedad será detectada tanto si la detectan ambas pruebas como si la detecta sólo una de ellas: $P(A = \text{sí})\, P(B = \text{sí}) + P(A = \text{sí})\, P(B = \text{no}) + P(A = \text{no})\, P(B = \text{sí}) = 0{,}90 \times 0{,}80 + 0{,}90 \times 0{,}20 + 0{,}10 \times 0{,}80 = 0{,}98$.

b. $P(B = \text{no})\; P(C = \text{no}) = 0{,}20 \times 0{,}30 = 0{,}06$.

c. $P(A = \text{sí})\; P(B = \text{sí})\; P(C = \text{sí}) = 0{,}90 \times 0{,}80 \times 0{,}70 = 0{,}504$.

d. Que al menos dos de las tres pruebas detecten la enfermedad significa que la detecten las tres o que la detecten sólo dos. Es decir,

$$P(A = \text{sí})\, P(B = \text{sí})\; P(C = \text{sí}) + P(A = \text{sí})\, P(B = \text{sí})\; P(C = \text{no}) + P(A = \text{sí})\, P(B = \text{no})\; P(C = \text{sí}) +$$
$$+ P(A = \text{no})\, P(B = \text{sí})\; P(C = \text{sí}) =$$
$$= (0{,}90 \times 0{,}80 \times 0{,}70) + (0{,}90 \times 0{,}80 \times 0{,}30) + (0{,}90 \times 0{,}20 \times 0{,}70) + (0{,}10 \times 0{,}80 \times 0{,}70) = 0{,}902.$$

2.11. Conocemos: $P(P_1) = P(P_2) = 0{,}50$; $P(A|P_1) = 1/4 = 0{,}25$; $P(A|P_2) = 1/5 = 0{,}20$. A partir de aquí se puede calcular la probabilidad de obtener un acierto: $P(A) = 0{,}50 \times 0{,}25 + 0{,}50 \times 0{,}20 = 0{,}225$. Por tanto:

a. $P(P_1|A) = P(P_1 \cap A)/P(A) = (0{,}50 \times 0{,}25)/0{,}225 = 0{,}56$.

b. $P(P_2|A) = P(P_2 \cap A)/P(A) = (0{,}50 \times 0{,}20)/0{,}225 = 0{,}44$.

2.12. *a*. El diagnóstico será correcto si el detector identifica correctamente tanto a una persona que miente ($M \cap m$) como a una persona que no miente ($nM \cap nm$):

- $P(M \cap m) = P(M)\; P(m \mid M) = 0{,}20 \times 0{,}90 = 0{,}18$.
- $P(nM \cap nm) = P(nM)\; P(nm \mid nM) = 0{,}80 \times 0{,}95 = 0{,}76$.

Por tanto, la probabilidad de diagnóstico correcto vale $0{,}18 + 0{,}76 = 0{,}94$.

b. Puesto que el detector indica que la persona miente (m), el diagnóstico sólo será correcto si la persona realmente está mintiendo (M). Hay que calcular: $P(M|m)$. Comencemos con lo que sabemos:

- Probabilidad de que un sujeto mienta: $P(M) = 0{,}20$.
- Probabilidad de que un sujeto no mienta: $P(nM) = 0{,}80$.
- $P(M \cap m) = 0{,}18$ (calculada en el apartado *a*).
- $P(nM \cap m) = 0{,}04$ (calculada en el apartado *a*).
- $P(m) = 0{,}18 + 0{,}04 = 0{,}22$.

Por tanto, $P(M|m) = P(M \cap m)/P(m) = 0{,}18/0{,}22 = 0{,}82$.

2.13. *a*. La prueba puede dar un resultado positivo (+) tanto si la enfermedad está presente (E) como si no (nE). Comencemos con lo que sabemos:

- Probabilidad de que la persona esté enferma: $P(E) = 0{,}30$.
- Probabilidad de la persona no esté enferma: $P(nE) = 0{,}70$.
- $P(E \cap +) = P(E)\,P(+|E) = 0{,}30 \times 0{,}90 = 0{,}27$.
- $P(nE \cap +) = P(nE)\,P(+|nE) = 0{,}70 \times 0{,}20 = 0{,}14$.

Por tanto, la probabilidad de obtener un resultado positivo vale $P(+) = 0{,}27 + 0{,}14 = 0{,}41$.

b. El diagnóstico será equivocado tanto si la prueba dice que una persona que está enferma (E) no lo está (−), como si dice que una persona que no está enferma (nE) lo está (+):

- $P(E \cap -) = P(E)\,P(-|E) = 0{,}30 \times 0{,}10 = 0{,}03$.
- $P(nE \cap +) = P(nE)\,P(+|nE) = 0{,}70 \times 0{,}20 = 0{,}14$.

Por tanto, la probabilidad de que el diagnóstico esté equivocado vale $0{,}03 + 0{,}14 = 0{,}17$.

c. $P(nE|+) = P(nE \cap +)/P(+) = 0{,}14/0{,}41 = 0{,}34$.

2.14. Hay que calcular: $P(S|C)$. Comencemos con lo que sabemos:

- Probabilidad de superar la prueba: $P(S) = 0{,}40$.
- Probabilidad de no superar la prueba: $P(nS) = 0{,}60$.
- $P(S \cap C) = 0{,}40 \times 0{,}80 = 0{,}32$.
- $P(nS \cap C) = 0{,}60 \times 0{,}05 = 0{,}03$.
- $P(C) = 0{,}32 + 0{,}03 = 0{,}35$.

Por tanto, $P(S|C) = P(S \cap C)/P(C) = 0{,}32/0{,}35 = 0{,}91$.

2.15. *a*. Hay que calcular $P(+)$. Comencemos con lo que sabemos:

- La probabilidad de ser chico (O): $P(O) = 0{,}48$.
- La probabilidad de ser chica (A): $P(A) = 0{,}52$.
- La probabilidad de tener 19 años o más entre los chicos: $P(+|O) = 0{,}15$.
- La probabilidad de tener 19 años o más entre las chicas: $P(+|A) = 0{,}05$.

Por tanto, $P(+) = 0{,}48 \times 0{,}15 + 0{,}52 \times 0{,}05 = 0{,}098$.

b. ¿$P(O \cap +) = P(O)\,P(+)$? No son independientes pues: $0{,}15 \neq 0{,}48 \times 0{,}098$.

c. Hay que calcular $P(A|+) = P(A \cap +)/P(+)$.

- $P(+) = 0{,}098$ (está calculado en el apartado *a*).
- $P(A \cap +) = P(A)\,P(+|A) = 0{,}52 \times 0{,}05 = 0{,}026$.

Por tanto, $P(A|+) = 0{,}026/0{,}098 = 0{,}265$.

2.16. *a*. Aproximadamente la mitad de los entrevistados que defraudan a Hacienda habrá obtenido cara en el primer lanzamiento de la moneda y habrá respondido "sí". La otra mitad habrá vuelto a tirar la moneda y la mitad de ellos habrá obtenido cruz y habrá respondido "sí". Por tanto, cabe esperar

que P valga la mitad de π más la cuarta parte de π. Es decir: $P = ½\pi + ¼\pi = 0{,}75\pi$. En consecuencia, utilizando esta estrategia podrá estimarse π mediante $P / 0{,}75$.

b. $P / 0{,}75 = 0{,}15 / 0{,}75 = 0{,}20$.

c. Lo primero que se está asumiendo es que los entrevistados responden con sinceridad cuando están seguros de que sus respuestas son completamente anónimas; este supuesto es razonable, pero no hay forma de constatar que efectivamente las cosas son así. También se está asumiendo que la proporción de caras y cruces que obtienen los entrevistados está igualada por el azar.

3

Análisis descriptivo de variables categóricas

Una vez recogidos los datos y preparados para el análisis, la primera tarea que conviene abordar (cualquiera que sea el tipo de análisis que finalmente se tenga intención de llevar a cabo para poder cubrir los objetivos de un estudio) es la de formarse una idea lo más exacta posible acerca de las características de cada variable.

Esto podría lograrse elaborando un listado de todos los valores, pues, cualquiera que sea la naturaleza de una variable, un listado de todos sus valores contiene toda la información disponible. Sin embargo, un listado de datos tiene poca (o quizá ninguna) utilidad. Lo realmente útil es poder organizar y resumir esos datos de alguna forma que aclare la situación y facilite la comprensión de lo que está ocurriendo. Ahora bien, como un resumen no es una descripción detallada de los datos, sólo podrá captar un aspecto parcial de los mismos. Por tanto, ya desde el principio, es importante conocer las herramientas estadísticas que permiten resumir los datos y acertar en la elección de aquellas que darán respuesta apropiada a las cuestiones que interese estudiar.

En este sentido, tanto si una variable es categórica como si es cuantitativa, obtener una descripción completa de la misma exige, por lo general, prestar atención a tres características claramente diferenciadas: el centro, la dispersión y la forma de la distribución (ver capítulo anterior). En este capítulo se presentan las herramientas estadísticas comúnmente utilizadas para describir variables categóricas; en el apéndice se describe un tipo particular de variables categóricas llamadas variables de *respuesta múltiple*.

Tablas de frecuencias

Una **tabla** o **distribución de frecuencias** es una forma particular de ordenar los datos basada en los valores concretos que adopta una variable categórica y en el número de veces que se repite cada valor. El objetivo de una tabla de frecuencias es organizar la información y, sobre todo, resumirla. La Tabla 3.1 muestra las frecuencias obtenidas al clasificar una muestra de $n = 200$ sujetos en la variable X = «tabaquismo». Se trata de una variable categórica que toma

tres valores: X_1 = «fumadores», X_2 = «exfumadores» y X_3 = «no fumadores» (el subíndice i se refiere a cada uno de los valores distintos que toma la variable; por tanto, i = 1, 2, 3).

La primera columna de la tabla recoge los tres valores de la variable. La segunda columna muestra las **frecuencias absolutas** (n_i), es decir, el número de veces que se repite cada valor. La tercera columna contiene las **frecuencias relativas** (P_i), las cuales se obtienen dividiendo las correspondientes frecuencias absolutas entre el número total de casos:

$$P_i = n_i / n \qquad [3.1]$$

Estas frecuencias indican la proporción de veces que se repite cada valor (también se les llama *proporciones*). Según tendremos ocasión de comprobar en próximos capítulos, las frecuencias relativas (las proporciones) son muy utilizadas para realizar comparaciones entre grupos.

Multiplicando por 100 las frecuencias relativas definidas en [3.1] se obtienen las **frecuencias porcentuales** ($\%_i$) que aparecen en la última columna de la tabla:

$$\%_i = 100 P_i \qquad [3.2]$$

Estas frecuencias indican el porcentaje de veces que se repite cada valor. Puesto que la información que ofrecen las frecuencias relativas y las porcentuales es idéntica, no es necesario incluir ambas en una misma tabla de frecuencias.

Tabla 3.1. Frecuencias de la variable *tabaquismo*

X *Tabaquismo*	n_i Frecuencias absolutas	P_i Frecuencias relativas	$\%_i$ Frecuencias porcentuales
Fumadores	70	0,35	35
Exfumadores	20	0,10	10
No fumadores	110	0,55	55
	200	1,00	100

La Tabla 3.2 ofrece las frecuencias de la variable *nivel de estudios*. Esta variable es categórica, al igual que la variable *tabaquismo*; pero, a diferencia de ésta, el *nivel de estudios* es una variable *ordinal* (sus categorías están cuantitativamente ordenadas). En estos casos es posible calcular un tipo particular de frecuencias llamadas *acumuladas*.

Tabla 3.2. Frecuencias de la variable *nivel de estudios*

X *Nivel de estudios*	n_i Frecuencias absolutas	P_i Frecuencias relativas	$\%_i$ Frecuencias porcentuales	na_i Frec. absol. acumuladas	Pa_i Frec. relat. acumuladas	$\%a_i$ Frec. porc. acumuladas
Sin estudios	5	0,01	1	5	0,01	1
Primarios	20	0,04	4	25	0,05	5
Secundarios	270	0,54	54	295	0,59	59
Medios	115	0,23	23	410	0,82	82
Superiores	90	0,18	18	500	1,00	100
	500	1,00	100			

La **frecuencia absoluta acumulada** (na_i) recoge el número de veces que se repite un valor más cualquier otro inferior a él. La **frecuencia relativa acumulada** (Pa_i) se obtiene dividiendo la frecuencia absoluta acumulada entre el número total de casos ($Pa_i = na_i / n$). Y la **frecuencia porcentual acumulada** ($\%a_i$) se obtiene multiplicando por 100 la frecuencia relativa acumulada ($\% a_i = 100Pa_i$).

Las frecuencias absolutas (n_i) constituyen el punto de referencia de una tabla de frecuencias: todas las demás frecuencias se calculan a partir de las absolutas. Por tanto, aunque una tabla concreta pueda incluir un tipo u otro de frecuencias en función de la información que interese destacar, es recomendable que las frecuencias absolutas (incluido el n total) siempre estén presentes.

Conviene señalar que una tabla de frecuencias no es una herramienta apropiada para describir variables cuantitativas. En general, si una variable cuantitativa se mide con suficiente precisión, no habrá valores que se repitan (o habrá muy pocos) y, en esas condiciones, una tabla de frecuencias será más un listado de casos que un resumen (no obstante, según veremos en el próximo capítulo, las frecuencias porcentuales acumuladas de una tabla de frecuencias constituyen la base de unos estadísticos muy útiles llamados *cuantiles*).

A pesar de su aparente simplicidad, una tabla o distribución de frecuencias contiene casi toda la información que encierra una variable categórica. Ofrece información útil sobre tres aspectos importantes. En primer lugar, indica qué valores toma la variable representada (en el ejemplo de la Tabla 3.1, la variable *tabaquismo* toma 3 valores: fumadores, exfumadores y no fumadores; en el ejemplo de la Tabla 3.2, la variable *nivel educativo* toma 5 valores: sin estudios, primarios, secundarios, medios y superiores). En segundo lugar, informa sobre qué valores son más frecuentes o se repiten más y qué valores son menos frecuentes o se repiten menos (por ejemplo, algo más de la mitad de los sujetos no fuma; un uno por ciento de los sujetos no tiene estudios, etc.). Por último, en el caso de variables categóricas ordinales, las frecuencias acumuladas indican cuántos sujetos alcanzan un determinado valor o están por encima de él (por ejemplo, aproximadamente el 60% de los sujetos no pasa de estudios obligatorios, es decir, secundarios).

Por tanto, de las tres propiedades o características a las que conviene prestar atención para describir apropiadamente una variable (centro, dispersión y forma de la distribución), una tabla de frecuencias ofrece información precisa sobre dos de ellas: el *centro* y la *forma* de la distribución.

El ***centro*** de la distribución es el valor con la frecuencia más alta (el valor que más se repite). Recibe el nombre de **moda**. En el ejemplo de la Tabla 3.1, la moda es «no fumadores»; en el de la Tabla 3.2, «secundarios». La moda entendida como centro de una distribución (es decir, como representante del resto de valores) tiene una capacidad descriptiva muy limitada, por lo que debe interpretarse con cautela: puede ocurrir que el valor que más se repite tenga una frecuencia baja; también puede ocurrir que haya más de una moda (categorías con la misma frecuencia) o que haya muy poca diferencia entre las dos categorías que más se repiten.

La ***forma*** de la distribución es visible a partir del tamaño de las frecuencias, que son las que indican dónde tienden a agruparse los valores y qué categorías tienen frecuencias pequeñas. Las frecuencias de la Tabla 3.1 indican que algo más de la mitad de los sujetos son no fumadores y que sólo el 10% son exfumadores. Las frecuencias de la Tabla 3.2 indican que algo más de la mitad de los sujetos tienen estudios secundarios y que sólo el 5% no pasa de estudios primarios. En relación con la forma de la distribución, es importante valorar la presencia de categorías con frecuencia nula o muy pequeña. Y si la variable es ordinal, hay que prestar atención a si las frecuencias se agrupan en torno al centro (simetría) o están desplaza-

das hacia uno de los extremos (asimetría). Por supuesto, un gráfico apropiado (ver siguiente apartado) puede ayudar a hacerse una buena idea de la forma de la distribución.

Sobre el tercer aspecto, la ***dispersión*** (grado de concentración o alejamiento de los valores en torno al centro de la distribución), la información de una tabla de frecuencias es poco precisa. Sabemos que la dispersión es mínima o nula cuando todas las frecuencias están concentradas en un mismo valor y máxima cuando están repartidas homogéneamente por todos ellos. Pero cuando no se da una de estas dos pautas no resulta nada fácil formarse una idea sobre el grado de dispersión. En estos casos puede utilizarse algún estadístico capaz de captar el grado de dispersión e informar de lo que está ocurriendo. Uno de estos estadísticos[1] es el *índice de variación cualitativa* (*IVC*):

$$IVC = \frac{\sum n_i n_{i'}}{I(I-1)(n/I)^2/2} \qquad [3.3]$$

(ver Mueller y Schuessler, 1961, págs. 177-179). El numerador de [3.3] recoge el número de diferencias observadas, es decir, el número de veces que un valor es distinto de cualquier otro (el subíndice i' indica un valor distinto de i). En el ejemplo sobre tabaquismo de la Tabla 3.1, el numerador toma el valor $70(20) + 70(110) + 20(110) = 11.300$. El denominador recoge el número máximo de posibles diferencias (I se refiere al número de categorías de la variable: $i = 1, 2, ..., I$). En el ejemplo de la Tabla 3.1, $3(3-1)(200/3)^2/2 = 13.333{,}33$. Y el cociente entre ambos resultados es el índice de variación cualitativa: $IVC = 11.300 / 13.333{,}33 = 0{,}85$. Aplicado a los datos de la Tabla 3.2, el *IVC* toma el valor 0,78.

El *IVC* toma valores comprendidos entre cero y uno. Cuando todas las frecuencias están concentradas en una sola categoría (dispersión mínima o nula), toma el valor cero; cuando las frecuencias están uniformemente repartidas entre todas las categorías (dispersión máxima) toma el valor uno. Por tanto, los valores obtenidos (0,85 y 0,74) están indicando, en ambos casos, un grado de dispersión medio-alto. Para facilitar la interpretación de este índice, el ejemplo que se ofrece más adelante sobre *tablas de frecuencias y gráficos de barras* incluye varias distribuciones con distinta dispersión acompañadas de sus respectivos *IVC*.

Gráficos para variables categóricas

Los informes sobre los resultados de un análisis descriptivo raramente se limitan a presentar la información numérica de una tabla de frecuencias; lo habitual es, más bien, acompañar esa información numérica con algún gráfico que permita formarse una impresión rápida de lo que está ocurriendo. Los gráficos más utilizados (también los más apropiados) con variables categóricas son los de *barras* y los de *sectores*. Son dos gráficos equivalentes en el sentido de que ofrecen la misma información, pero su aspecto es muy distinto.

Un **gráfico de barras** (ver Figura 3.1) se construye sobre el plano definido por dos ejes cartesianos: en el eje horizontal se colocan los valores de la variable; en el vertical, las frecuencias; y sobre cada valor se levanta una barra de altura proporcional a su frecuencia (la anchura de las barras no es relevante, pero todas ellas han de tener la misma). Las barras se

[1] Existen otros índices para valorar la dispersión de una variable categórica (concentración, entropía, etc.), pero no es común incluirlos en un informe descriptivo (se utilizan más para estudiar la relación entre variables) y, por tanto, no serán tratados aquí (ver, por ejemplo Agresti, 2002, págs. 24-25; o Wickens, 1989, pág. 130).

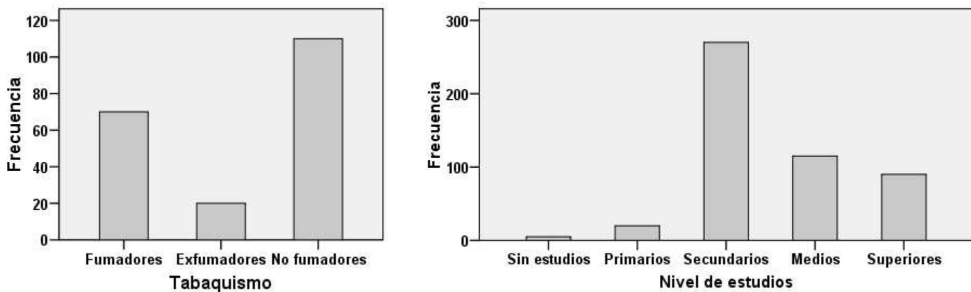
representan separadas para resaltar la idea de que corresponden a distintos valores de la variable. La Figura 3.1 muestra los gráficos de barras correspondientes a las frecuencias de las Tablas 3.1 y 3.2. Elegir uno u otro tipo de frecuencias (absolutas, relativas o porcentuales) afecta a los números que aparecen en el eje vertical, pero no a la forma del gráfico.

Cuando la variable es categórica *ordinal* (variable con categorías cuantitativamente ordenadas), también pueden representarse las *frecuencias acumuladas* colocando el valor menor a la izquierda y el mayor a la derecha (no obstante, dado que una variable ordinal suele estar midiendo una característica continua, lo habitual es representar este tipo de variables mediante histogramas; ver siguiente capítulo).

Cuando se construyen gráficos de barras hay que tomar algunas precauciones para no distorsionar la información que se está ofreciendo. En primer lugar, debe evitarse cortar el eje vertical, pues, si no se representa toda su altura, la diferencia entre las barras puede resultar engañosa (pequeñas diferencias pueden parecer muy grandes). En segundo lugar, es muy desaconsejable sustituir las barras por figuras o dibujos de aquello que se quiere representar (por ejemplo, figuras humanas para representar el número o porcentaje de personas, dibujos de coches para representar el número de ventas, etc.); la razón de esto es que los dibujos correspondientes a las frecuencias más altas no sólo son más altos sino, además, más anchos, con lo que el área que ocupan (y la consiguiente impresión visual que producen) es sensiblemente mayor que su correspondiente frecuencia.

Según tendremos ocasión de comprobar, los gráficos de barras no sólo son útiles para representar variables categóricas (es decir, variables medidas con una escala nominal u ordinal); también lo son para representar variables cuantitativas discretas cuando éstas toman sólo unos pocos valores (número de hijos, número de aciertos en una prueba de 10 preguntas, etc.).

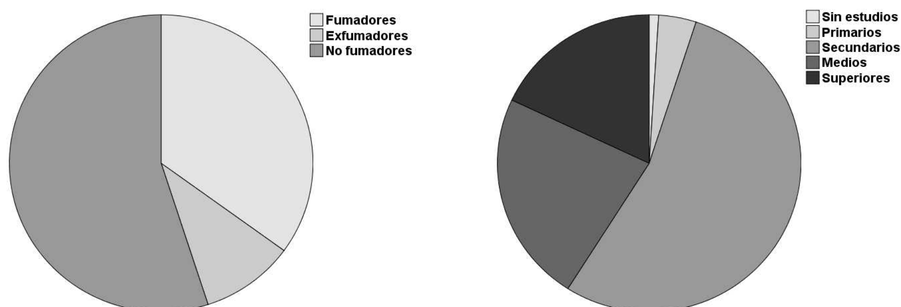
Figura 3.1. Gráficos de barras de las variables *tabaquismo* (izquierda) y *nivel de estudios* (derecha)



Una herramienta alternativa para representar variables categóricas es el **gráfico de sectores** (también llamado de *tarta*, de *quesitos*, etc.; ver Figura 3.2). Se construye dividiendo un círculo en tantos sectores como valores distintos (categorías) toma la variable representada y asignando a cada valor un sector de tamaño proporcional a su frecuencia. El tamaño de cada sector se obtiene asignándole el ángulo resultante de multiplicar por 360 la correspondiente frecuencia relativa ($360P_i$), pero la forma del gráfico no se ve afectada por elegir uno u otro tipo de frecuencias (absolutas, relativas o porcentuales).

En general, un gráfico de sectores no permite comparar el tamaño de los sectores con la misma facilidad que lo hace un gráfico de barras. Y pierde eficacia cuando la variable tiene muchas categorías. No obstante, tiene una interesante ventaja: es posible, en caso de que se considere conveniente, destacar un sector separándolo del resto.

Figura 3.2. Gráficos de sectores de las variables *tabaquismo* (izquierda) y *nivel de estudios* (derecha)



Análisis descriptivo de variables categóricas con SPSS

Las herramientas descriptivas estudiadas en este capítulo (las tablas de frecuencias y los gráficos de barras y sectores) están incluidas, entre otros, en el procedimiento **Frecuencias**. Este procedimiento también permite obtener algunos de los estadísticos descriptivos más conocidos y utilizados (posición, tendencia central, dispersión y forma de la distribución), pero éstos los estudiaremos en el próximo capítulo.

Al cuadro de diálogo **Frecuencias** se accede con la opción **Estadísticos descriptivos > Frecuencias** del menú **Analizar**. Dentro del cuadro de diálogo, la opción **Mostrar tablas de frecuencias**, que se encuentra activa por defecto, permite decidir si se desea o no obtener la distribución de frecuencias. Esta opción puede desactivarse si sólo interesa ver algún gráfico o algún estadístico descriptivo. Si se desactiva esta opción y no se selecciona ninguna otra, los resultados únicamente muestran el número total de casos válidos del archivo y el número de casos con valor perdido.

El botón **Gráficos** conduce al subcuadro de diálogo **Frecuencias: Gráficos**, el cual permite obtener los gráficos de barras y sectores estudiados en este capítulo[2]. Las opciones del recuadro **Valores del gráfico** permiten elegir el tipo de frecuencia que se desea representar: absolutas (opción **Frecuencias**) o porcentuales (opción **Porcentajes**).

Ejemplo. Tablas de frecuencias y gráficos de barras con SPSS

Este ejemplo muestra cómo obtener una tabla o distribución de frecuencias con el procedimiento **Frecuencias** del SPSS. También muestra cómo obtener gráficos de barras. El ejemplo se basa en el archivo *Datos de empleados* (que se encuentra en la misma carpeta en la que está instalado el SPSS), el cual contiene una muestra de 474 empleados de banca en los que se han registrado variables como el sexo, la fecha de nacimiento, el nivel educativo, el salario, la categoría laboral, etc. Para obtener tablas de frecuencias y gráficos de barras:

[2] El menú **Gráficos** de la barra de menús principal también ofrece diferentes opciones para obtener estos mismos gráficos más otros llamados *interactivos* que poseen mejores prestaciones y más posibilidades estéticas.

➤ Seleccionar las variables *catlab* (categoría laboral) y *educ* (nivel educativo) y trasladarlas a la lista **Variables**.

➤ Pulsar el botón **Gráficos** para acceder al subcuadro de diálogo **Frecuencias: Gráficos** y marcar la opción **Gráficos de barras**; pulsar el botón **Continuar** para volver al cuadro de diálogo principal.

Aceptando estas elecciones, el *Visor* ofrece los resultados que muestran las Tablas 3.3 y 3.4 y los gráficos de barras que muestra la Figura 3.3. Las tablas informan sobre: (1) los valores que toma la variable (*válidos*); por supuesto, si la Tabla 3.4 no muestra valores como el 9, el 10 y el 11 es porque la variable *nivel educativo* no toma esos valores; (2) la frecuencia absoluta de cada valor (*frecuencia*); (3) la frecuencia porcentual calculada sobre el número total de casos del archivo (*porcentaje*); (4) la frecuencia porcentual calculada sobre el número de casos válidos, es decir, el número total de casos menos el número de casos con valor perdido (*porcentaje válido*); y (5) la frecuencia porcentual acumulada (*porcentaje acumulado*) calculada a partir del porcentaje válido (el *porcentaje* y el *porcentaje válido* son iguales cuando no existen valores perdidos). La última línea ofrece el número total de casos.

El *nivel educativo* no es una variable categórica, sino una variable cuantitativa discreta (años de formación académica), pero se ha incluido deliberadamente en el ejemplo para señalar una idea que no debe pasar inadvertida: aunque las tablas de frecuencias son herramientas especialmente útiles para describir variables categóricas (y así es como lo hemos hecho en los ejemplos correspondientes a las Tablas 3.1 y 3.2), también son útiles para describir variables cuantitativas discretas que toman sólo unos pocos valores distintos; lógicamente, con variables que toman muchos valores distintos, la tabla de frecuencias pierde eficacia informativa.

La información numérica de las Tablas 3.3 y 3.4 está representada en los gráficos de barras de la Figura 3.3. En ambos gráficos puede apreciarse con claridad la diferencia existente entre las frecuencias de las distintas categorías.

Con la información que proporcionan las tablas de frecuencias y los gráficos de barras es posible formarse una idea bastante precisa sobre el centro de las distribuciones, sobre su grado de dispersión y sobre su forma. Por lo que se refiere al ***centro*** de las distribuciones, la moda de *categoría laboral* (la frecuencia más alta o valor que más se repite) es «administrativo»; la moda de *nivel educativo* es 12.

En ambas variables ocurre que un alto porcentaje de casos se ubica en la categoría modal, pero esto es mucho más evidente en *categoría laboral* que en *nivel educativo* (76,6% frente a 40,1%), lo cual revela diferente grado de ***dispersión***; de hecho, el índice de variación cualitativa *IVC* (no lo calcula el SPSS) vale 0,57 para *categoría laboral* y 0,83 para *nivel educativo*. La Figura 3.4 muestra algunos gráficos de barras que pueden ayudar a comprender el significado de este índice. Estos gráficos están ordenados de menor a mayor dispersión: el *IVC* que acompaña a cada gráfico indica que la dispersión es baja cuando los casos tienden a acumularse en una sola categoría, y que va aumentando paulatinamente a medida que las barras (frecuencias) se van igualando.

Por último, la ***forma*** de la distribución permite constatar, básicamente, si existen categorías que tienden a aglutinar muchos casos y si existen categorías con frecuencias muy pequeñas o nulas. Así, en la tabla correspondiente a *categoría laboral* se observa que, mientras el 76,6% de los casos son administrativos, sólo el 5,6% son agentes de seguridad. Y en la de *nivel educativo* se observan dos valores con una frecuencia alta (12, 15), tres con una frecuencia intermedia (8, 16, 19) y cuatro con una frecuencia baja (14, 17, 18, 20, 21). Además, como los valores de la variable *nivel educativo* están cuantitativamente ordenados, tiene sentido

prestar atención a las frecuencias porcentuales acumuladas: el porcentaje acumulado de la Tabla 3.4 indica que algo más de la mitad de los casos (51,3%) tienen 12 años de formación o menos (aproximadamente el equivalente a enseñanza obligatoria) y que menos del 10% de los casos tiene más de 17 años de formación (aproximadamente el equivalente a formación superior).

Tabla 3.3. Frecuencias de la variable *catlab* (categoría laboral)

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	Administrativo	363	76,6	76,6	76,6
	Seguridad	27	5,7	5,7	82,3
	Directivo	84	17,7	17,7	100,0
	Total	474	100,0	100,0	

Tabla 3.4. Frecuencias de la variable *educ* (nivel educativo)

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	8	53	11,2	11,2	11,2
	12	190	40,1	40,1	51,3
	14	6	1,3	1,3	52,5
	15	116	24,5	24,5	77,0
	16	59	12,4	12,4	89,5
	17	11	2,3	2,3	91,8
	18	9	1,9	1,9	93,7
	19	27	5,7	5,7	99,4
	20	2	,4	,4	99,8
	21	1	,2	,2	100,0
	Total	474	100,0	100,0	

Figura 3.3. Gráficos de barras de las variables *categoría laboral* (izquierda) y *nivel educativo* (derecha)

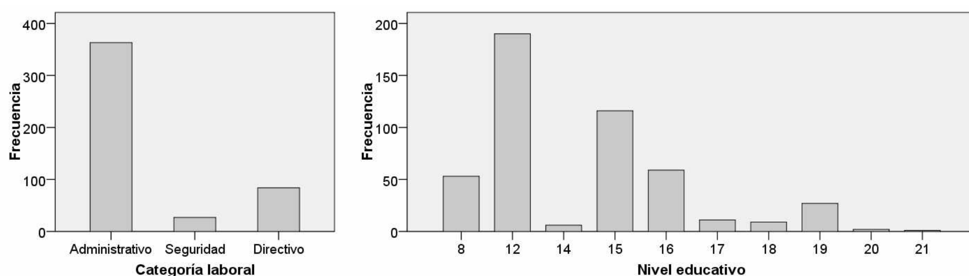
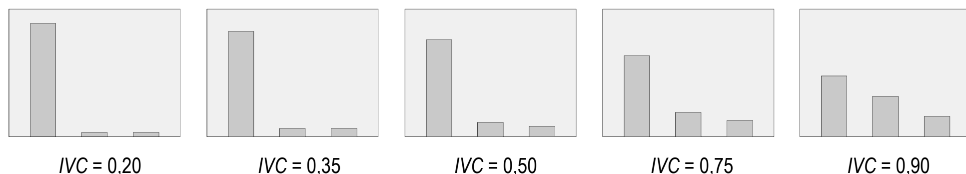


Figura 3.4. Gráficos de barras con diferente grado de dispersión (*IVC* = índice de variación cualitativa)



IVC = 0,20 *IVC* = 0,35 *IVC* = 0,50 *IVC* = 0,75 *IVC* = 0,90

Variables dicotómicas

Las variables dicotómicas son variables categóricas que toman sólo dos valores: acierto-error, verdadero-falso, recuperados-no recuperados, a favor-en contra, funciona-no funciona, etc. Muchas de estas variables se obtienen registrando la *presencia-ausencia* de algún evento (por ejemplo, la presencia-ausencia de un tratamiento, o de un síntoma); otras muchas se obtienen dicotomizando variables cuantitativas (por ejemplo, aprobados-suspensos, hipertensos-no hipertensos, etc.). Al trabajar con este tipo de variables es costumbre codificar con "unos" la presencia de la característica, los aciertos, los recuperados, etc., y con "ceros" la ausencia de la característica, los errores, los no recuperados, etc. (ver Tabla 3.5).

Centrar la atención en este tipo de variables está justificado por el hecho de tratarse de variables con una destacada presencia en diferentes áreas de conocimiento. La Tabla 3.5 recoge el número de sujetos que padecen trastorno depresivo en una muestra aleatoria de hombres mayores de 40 años. ¿Qué tipo de información podría destacarse de esta tabla? De acuerdo con lo ya estudiado en apartados anteriores a propósito de la descripción de variables categóricas, habría que decir que la categoría modal (el centro de la distribución) es X_0 = «no» y que la dispersión es media-alta (*IVC* = 0,64), con una evidente desproporción entre las frecuencias de las dos categorías (20% frente a 80%).

Éste es el tipo de información a la que se dirige la atención cuando únicamente se tienen en cuenta los datos muestrales obtenidos sin otro tipo de consideración adicional. Pero ocurre que, con frecuencia, se trabaja con poblaciones de las que se conocen (o se asume que se conocen) algunas de sus características. Así, por ejemplo, a partir de estudios previos es posible conocer con cierta precisión la proporción de fumadores de una determinada población, o el porcentaje de éxito de un determinado tratamiento, o la prevalencia de una determinada enfermedad, etc. Imaginemos, por ejemplo, que los datos de la Tabla 3.5 se han obtenido de una población de la que se sabe que la prevalencia de trastorno depresivo es del 15%. En este escenario, la probabilidad teórica de encontrar sujetos con trastorno depresivo (π_1) vale 0,15 y el conocimiento de este dato puede ayudar a entender mejor los resultados obtenidos en una muestra concreta. De hecho, el conocimiento de este dato permite utilizar la *distribución binomial* para conocer las probabilidades asociadas a cada posible resultado muestral.

Tabla 3.5. Frecuencias de la variable *depresión*

X = *Depresión*	n_i	P_i	π_i
1 = Sí	4	0,2	0,15
0 = No	16	0,8	0,85
	20	1	1

La distribución binomial

Un ensayo que sólo puede tomar uno de dos resultados posibles mutuamente exclusivos (lanzar una moneda –cara, cruz–, observar el resultado de una respuesta –acierto, error–, constatar el resultado de un tratamiento –recuperación, no recuperación–, etc.) se conoce como **ensayo de Bernoulli** (en recuerdo del matemático del mismo nombre). Aunque es una práctica bas-

tante generalizada llamar *éxito* a uno de estos resultados y *fracaso* al otro (sin que esto tenga necesariamente relación con la naturaleza positiva o negativa del resultado), en muchos contextos es preferible utilizar una terminología que capte con mayor precisión el significado del ensayo, como presencia-ausencia, acierto-error, sucede-no sucede, etc.

La **distribución binomial** permite conocer la probabilidad asociada al *número de éxitos* (o de *fracasos*) obtenidos en un conjunto de ensayos de Bernoulli, es decir, la probabilidad de obtener un determinado número de caras en un conjunto de lanzamientos de una moneda, un determinado número de aciertos en un conjunto de respuestas, un determinado número de recuperaciones en un conjunto de pacientes tratados, etc. Por tanto, la distribución binomial sirve para trabajar con variables dicotómicas. Pero no exactamente con los dos valores que toma una variable dicotómica X (uno-cero; éxito-fracaso), sino con el *número de éxitos* (n_1) o el *número de fracasos* (n_0) observados[3] en un conjunto de n ensayos, registros o réplicas de una variable dicotómica. Matemáticamente, la distribución binomial se define así:

$$P(n_1) \;=\; \binom{n}{n_1}\pi_1^{n_1}(1-\pi_1)^{n-n_1} \;=\; \frac{n!}{n_1!\,(n-n_1)!}\;\pi_1^{n_1}\;(1-\pi_1)^{n-n_1} \qquad [3.4]$$

donde n se refiere al número de ensayos; n_1 es el valor concreto de la variable *número de éxitos* (caras, recuperaciones) cuya probabilidad se desea conocer; y π_1 es la probabilidad teórica de éxito[4]. Esta ecuación permite calcular la probabilidad de que la variable n_1 = «número de éxitos» tome cada uno de sus posibles valores: 0, 1, 2, ..., n (por supuesto, lo mismo vale decir de la variable n_0 = «número de fracasos» si n_1 y π_1 se sustituyen por n_0 y π_0). El único requisito para obtener estas probabilidades es que los ensayos sean independientes entre sí (es decir, que el resultado de un ensayo no condicione el resultado de otro ensayo) o, dicho de otro modo, que las probabilidades de éxito y de fracaso, π_1 y π_0, permanezcan constantes en cada ensayo. Cuando se da esta circunstancia decimos que la variable n_1 = «número de éxitos» se distribuye binomialmente con parámetros n y π_1, lo cual se representa mediante[5]

$$n_1 \;\sim\; B(n, \pi_1) \qquad [3.5]$$

Aplicando la ecuación [3.4] al ejemplo de la Tabla 3.5, donde $n = 20$ y $\pi_{\text{depresión = sí}} = \pi_1 = 0{,}15$, la probabilidad de encontrar 4 sujetos con trastorno depresivo vale

$$P\,(n_1 = 4) \;=\; \frac{20!}{4!\,(20-4)!}\,0{,}15^4(1-0{,}15)^{20-4} \;=\; 0{,}182$$

[3] Recordemos (ver Capítulo 2) que en una situación como la descrita en la Tabla 3.5 no sólo existe la variable X = «presencia-ausencia de trastorno depresivo». Además de esa variable, cuyos valores son fijos (1 = «sí»; 0 = «no»), existen otras dos; a saber, n_1 = «número de pacientes con trastorno depresivo» y n_0 = «número de pacientes sin trastorno depresivo». Son variables porque sus valores cambian (varían) dependiendo de la muestra concreta en la que se obtienen. Recordemos que a estas variables cuyos valores dependen del muestreo aleatorio se les llama *variables aleatorias* y, para caracterizarlas, hay que prestar atención a los tres aspectos en los que venimos insistiendo: centro, dispersión y forma de la distribución.

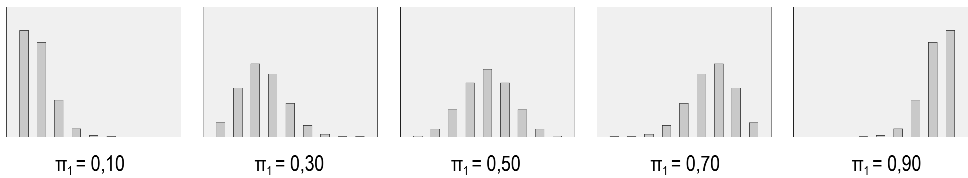
[4] En un ensayo de Bernoulli (con posibles resultados $X = 1$ = «éxito» y $X = 0$ = «fracaso») se está asumiendo $P(X=1) = \pi_1$ y $P(X=0) = \pi_0$. En dos ensayos: $P(X=1, X=1) = \pi_1\pi_1$; $P(X=1, X=0) = \pi_1\pi_0$; $P(X=0, X=1) = \pi_0\pi_1$; $P(X=0, X=0) = \pi_0\pi_0$. Etcétera. A partir de aquí es fácil deducir la función [3.4] (ver, por ejemplo, Amón, 1984, págs. 87-90).

[5] Puesto que n y π_1 pueden tomar distintos valores, en realidad no existe una única distribución binomial, sino una familia de distribuciones binomiales (tantas como valores distintos puedan tomar n y π_1; ver Figura 3.5), todas las cuales se ajustan a la misma regla.

Por supuesto, para conocer las probabilidades de una distribución binomial no es necesario efectuar cálculos a mano. Esas probabilidades están ya tabuladas y pueden encontrarse en cualquier manual de estadística o de análisis de datos (ver siguiente apartado). Por otro lado, el SPSS, al igual que otros programas estadísticos, incluye funciones que permiten calcular estas probabilidades (ver siguiente ejemplo).

La Figura 3.5 muestra la forma de varias distribuciones binomiales para $n=8$ y diferentes valores de π_1. En el eje horizontal están representados los valores de la variable n_1 = «número de éxitos» (de 0 a 8); en el vertical, la probabilidad asociada a cada uno de esos valores. Estos gráficos muestran con claridad dos cosas: (1) la distribución binomial es simétrica cuando π_1 vale 0,5 (gráfico del centro) y (2) se va volviendo más y más asimétrica a medida que el valor de π_1 se va alejando de 0,5.

Figura 3.5. Distribuciones binomiales para n = 8 y diferentes valores de π_1



Recordemos ahora que para describir correctamente una variable hay que prestar atención a tres propiedades básicas de su distribución: centro, dispersión y forma. Acabamos de ver que la **forma** de la distribución de la variable *número de éxitos* la ofrece la distribución binomial. Pero, ¿qué puede decirse del centro y de la dispersión? El **centro** de la distribución de una variable aleatoria es el valor que cabe esperar encontrar con mayor probabilidad. Recordemos (ver, en el Capítulo 2, el apartado *Variables aleatorias*) que a ese valor se le llama *valor esperado* (E). Con variables discretas, el valor esperado se define de la siguiente manera:

$$E(X) \;=\; \sum_i X_i\, f(X_i) \qquad [3.6]$$

Es decir, sumando los productos entre cada valor de la variable, X_i, y su correspondiente frecuencia relativa $f(X_i)$. Como n_1 = «número de éxitos» es una variable resultante de sumar n ensayos de Bernoulli (unos y ceros), tendremos[6]

$$E(n_1) \;=\; \mu_{n_1} \;=\; n\pi_1 \qquad [3.7]$$

Aplicando esta ecuación a los datos de la Tabla 3.5, donde n vale 20 y π_1 vale 0,15, el centro o valor esperado de la variable *número de pacientes con trastorno depresivo* toma el valor $E(n_1) = 20(0{,}15) = 3$. Esto significa que, al seleccionar una muestra aleatoria de 20 sujetos de una población donde la prevalencia de trastorno depresivo es del 15%, el número de sujetos con trastorno depresivo que cabe esperar encontrar con mayor probabilidad es 3.

---

[6] En efecto, como n_1 es la suma de n variables (n ensayos de Bernoulli) y el valor esperado de una suma de variables es igual a la suma de sus valores esperados, se verifica que

$$E(n_1) = E(X_1 + X_2 + \cdots + X_n) = E(X_1) + E(X_2) + \cdots + E(X_n)$$

Además, por la ecuación [3.6] y teniendo en cuenta que las variables X_i son dicotómicas (sólo toman unos y ceros), sabemos que $E(X_1) = 1(\pi_1) + 0(\pi_0) = \pi_1$; $E(X_2) = 1(\pi_1) + 0(\pi_0) = \pi_1$; etc. Consecuentemente,

$$E(n_1) = \pi_1 + \pi_1 + \cdots + \pi_1 = n\pi_1 .$$

Por último, el grado de **dispersión** del *número de éxitos* puede cuantificarse mediante un estadístico llamado *varianza* que se define de la siguiente manera:

$$V(X) \;=\; \sigma_X^2 \;=\; E(X^2) - [E(X)]^2 \qquad [3.8]$$

Y como n_1 = «número de éxitos» es una variable resultante de sumar n ensayos de Bernoulli, tendremos[7]

$$V(n_1) \;=\; \sigma_{n_1}^2 \;=\; n\,\pi_1(1-\pi_1) \qquad [3.9]$$

En el ejemplo propuesto en la Tabla 3.5 (donde $n = 20$ y $\pi_1 = 0{,}15$), la varianza de la variable *número de sujetos con trastorno depresivo* vale $20(0{,}15)(1-0{,}15) = 2{,}55$. En los próximos capítulos tendremos ocasión de familiarizarnos con el significado de estos estadísticos.

Ejemplo. Variables dicotómicas y distribución binomial

Supongamos que el servicio de psicoterapia de un determinado hospital consigue recuperaciones aceptables con el 70% de los pacientes que padecen trastorno depresivo. ¿Cuál es la probabilidad de encontrar más de 11 recuperaciones en una muestra aleatoria de 15 pacientes tratados en ese hospital?

Se tienen $n = 15$ observaciones de una variable dicotómica (1 = «recuperados», 0 = «no recuperados») y se asume que la probabilidad de recuperación vale $\pi_1 = 0{,}70$ en cada observación. La probabilidad de obtener más de 11 recuperaciones ($n_1 > 11$) equivale a la probabilidad de obtener $n_1 = 12$, $n_1 = 13$, $n_1 = 14$, o $n_1 = 15$ recuperaciones. Es decir,

$$P(n_1 > 11) \;=\; P(n_1 = 12) + P(n_1 = 13) + P(n_1 = 14) + P(n_1 = 15)$$

Aplicando ahora [3.4] con $n = 15$ y $\pi_1 = 0{,}70$ se obtienen las siguientes probabilidades:

$$P(n_1 = 12) \;=\; \frac{15!}{12!\,(15-12)!}\,0{,}70^{12}\,(1-0{,}70)^{15-12} \;=\; 0{,}1700$$

$$P(n_1 = 13) \;=\; \frac{15!}{13!\,(15-13)!}\,0{,}70^{13}\,(1-0{,}70)^{15-13} \;=\; 0{,}0916$$

$$P(n_1 = 14) \;=\; \frac{15!}{14!\,(15-14)!}\,0{,}70^{14}\,(1-0{,}70)^{15-14} \;=\; 0{,}0305$$

$$P(n_1 = 15) \;=\; \frac{15!}{15!\,(15-15)!}\,0{,}70^{15}\,(1-0{,}70)^{15-15} \;=\; 0{,}0047$$

---

[7] En efecto, como n_1 es la suma de n variables independientes (n ensayos de Bernoulli) y la varianza de una suma de variables independientes es igual a la suma de sus varianzas, se verifica:

$$V(n_1) = \sigma_{n_1}^2 = \sigma_{X_1+X_2+\cdots+X_n}^2 = \sigma_{X_1}^2 + \sigma_{X_2}^2 + \cdots + \sigma_{X_n}^2$$

Además, sabemos que $E(X_1^2) = E(X_2^2) = \cdots = E(X_n^2) = 1^2(\pi_1) + 0^2(\pi_0) = \pi_1$. Por tanto, de acuerdo con [3.8], $\sigma_{X_i}^2 = \pi_1 - \pi_1^2$. En consecuencia,

$$V(n_1) = (\pi_1 - \pi_1^2) + (\pi_1 - \pi_1^2) + \cdots + (\pi_1 - \pi_1^2) = n(\pi_1 - \pi_1^2) = n\pi_1(1-\pi_1).$$

Por tanto, la probabilidad de encontrar más de 11 recuperaciones (en $n = 15$ ensayos con probabilidad de recuperación $\pi_1 = 0{,}70$) vale: $0{,}1700 + 0{,}0916 + 0{,}0305 + 0{,}0047 = 0{,}297$.

Supongamos ahora que un laboratorio afirma que un determinado fármaco produce efectos secundarios en 10 de cada 100 pacientes. Si se eligen al azar 30 pacientes que han tomado el fármaco: (1) ¿cuántos cabe esperar que presenten efectos secundarios?; (2) ¿cuál es la probabilidad de encontrar un solo paciente con efectos secundarios; (3) ¿cuál es la probabilidad de encontrar al menos 4 pacientes con efectos secundarios?

La variable n_1 = «número de pacientes que presentan efectos secundarios» se distribuye binomialmente con parámetros $n = 30$ y $\pi_1 = 0{,}10$. Por tanto:

(1) Valor esperado: $E(n_1) = n\pi_1 = 30\,(0{,}10) = 3$ pacientes.

(2) $P(n_1 = 1) = \dfrac{30!}{1!\,(30-1)!}\,0{,}10^1\,(0{,}90)^{30-1} = 0{,}141.$

(3) La probabilidad de encontrar 4 o más pacientes con efectos secundarios puede obtenerse restando a 1 la probabilidad de encontrar menos de cuatro pacientes; por tanto, es necesario conocer la probabilidad de $n_1 = 0$, $n_1 = 1$, $n_1 = 2$ y $n_1 = 3$. Aplicando la ecuación [3.4] se obtiene: $P(n_1 = 0) = 0{,}042$; $P(n_1 = 1) = 0{,}141$; $P(n_1 = 2) = 0{,}228$; $P(n_3 = 3) = 0{,}236$. Consecuentemente, $P(n_1 < 4) = 0{,}042 + 0{,}141 + 0{,}228 + 0{,}236 = 0{,}647$. Restando ahora de 1 esta probabilidad se obtiene $P(n_1 \geq 4) = 1 - P(n_1 < 4) = 1 - 0{,}647 = 0{,}353$.

Utilizando la tabla de la distribución binomial y algunas de las funciones que incluye el SPSS pueden obtenerse estos mismos resultados.

Tabla de la distribución binomial

Los cálculos realizados en el ejemplo anterior son innecesarios si se utiliza la tabla de la distribución binomial (ver la Tabla B del apéndice final). La tabla ofrece las probabilidades asociadas a n_1 para algunos valores de n y π_1 (el subíndice 1 se refiere a la categoría de referencia; normalmente, *éxito*). Los valores de n van de 1 a 20 y definen los 20 subgrupos de filas que contiene la tabla. Los valores de π_1 van de 0,10 a 0,90 y definen las columnas de la tabla. Los valores del interior de la tabla son las probabilidades binomiales. Debe tenerse muy presente que las probabilidades que ofrece la tabla son probabilidades *acumuladas*.

Así, con $n = 15$ y $\pi_1 = 0{,}70$, la tabla asocia a $n_1 = 11$ una probabilidad de 0,703; este valor indica la probabilidad de encontrar 11 éxitos o menos (probabilidad acumulada):

$$P(n_1 \leq 11) = F(11) = 0{,}703$$

Restando de 1 la probabilidad acumulada hasta el valor 11 se obtiene la probabilidad de encontrar más de 11 éxitos:

$$P(n_1 > 11) = 1 - F(11) = 1 - 0{,}703 = 0{,}297$$

Puede comprobarse que este valor (0,297) coincide con el ya obtenido en el ejemplo anterior aplicando la ecuación [3.4].

Para conocer la probabilidad asociada a un valor concreto, basta con buscar su probabilidad acumulada y restarle la acumulada hasta el valor inmediatamente anterior a él. Así, por ejemplo, para calcular la probabilidad de obtener $n_1 = 4$ éxitos en $n = 10$ ensayos cuando la probabilidad de éxito vale $\pi_1 = 0{,}20$, hay que buscar la probabilidad acumulada hasta el valor 4 y restarle la acumulada hasta el valor 3:

$$P(n_1 \leq 4) \;=\; F(4) \;=\; 0{,}967$$

$$P(n_1 \leq 3) \;=\; F(3) \;=\; 0{,}879$$

Por tanto, $P(n_1 = 4) \;=\; F(4) - F(3) \;=\; 0{,}967 - 0{,}879 \;=\; 0{,}088$.

La distribución binomial con SPSS

La tabla de la distribución binomial tiene sus limitaciones: únicamente incluye algunos valores de n y π_1. Para otros valores de n o π_1 es necesario utilizar otra estrategia. El SPSS incluye varias funciones relacionadas con la distribución binomial, todas las cuales se encuentran en la opción **Calcular** del menú **Transformar**.

La función **CDF.BINOM** $(\boldsymbol{n_1, n, \pi_1})$ calcula la probabilidad acumulada hasta el valor n_1 en n ensayos cuando la probabilidad de éxito es π_1. Por tanto,

CDF.BINOM (11, 15, 0.70) = 0.703

es la probabilidad de obtener 11 éxitos o menos ($n_1 \leq 11$) en 15 ensayos ($n = 15$) cuando la probabilidad de éxito vale 0,70 ($\pi_1 = 0{,}70$). Restando de 1 este resultado se obtiene la probabilidad de obtener más de 11 éxitos: 1 − CDF.BINOM (11, 15, 0.70) = 0.297.

La función **PDF.BINOM**$(\boldsymbol{n_1, n, \pi_1})$ permite obtener la probabilidad individual asociada a n_1 éxitos en n ensayos con probabilidad de éxito π_1. Esto equivale a aplicar la ecuación [3.4]. Así, por ejemplo, la probabilidad de obtener un éxito ($n_1 = 1$) en 30 ensayos ($n = 30$) con una probabilidad de éxito de 0,10 ($\pi_1 = 0{,}10$) puede obtenerse mediante

PDF.BINOM(1, 30, 0.10) = 0.141

Este valor coincide con el obtenido en el ejemplo anterior aplicando [3.4]. No debe olvidarse que el separador decimal que debe utilizarse en las expresiones numéricas del SPSS es el punto (como en una calculadora), no la coma (como se hace al escribir en español).

Variables politómicas

Una variable politómica es una variable categórica que toma más de dos valores. Los ejemplos de las Tablas 3.1 a la 3.4 y el de la Tabla 3.6 recogen variables politómicas. Algunas de estas variables, como el *tabaquismo*, se obtienen aplicando una escala de medida *nominal*; otras, como el *nivel de estudios*, se obtienen aplicando una escala de medida *ordinal*. El tipo de tratamiento (A, B, control), la gravedad de un trastorno (leve, moderado, grave), la actitud hacia sujetos, objetos o conductas (favorable, desfavorable o indiferente), la raza, las preferencias políticas, el estado civil, la clase social, el lugar de residencia, la ocupación laboral, etc., son algunos ejemplos de variables politómicas.

Tabla 3.6. Frecuencias de la variable *tabaquismo*

X = *Tabaquismo*	n_i	P_i	π_i
1 = Fumadores	17	0,34	0,30
2 = Exfumadores	5	0,10	0,05
3 = No fumadores	28	0,56	0,65
	50	1,00	1,00

Las herramientas estadísticas disponibles para describir este tipo de variables ya se han expuesto. Pero, al igual que ocurre con las variables dicotómicas, los resultados de una variable politómica también pueden obtenerse al muestrear poblaciones cuyas características se conocen o se asumen conocidas. Imaginemos, por ejemplo, que los datos de la Tabla 3.6 se han obtenido de una población de la que se sabe, por estudios previos, que está compuesta por un 30% de fumadores, un 5% de exfumadores y un 65% de no fumadores. En este escenario, las probabilidades teóricas (π_i) de encontrar fumadores, exfumadores y no fumadores valen, respectivamente, 0,30, 0,05 y 0,65, y el conocimiento de estas probabilidades puede ayudar a entender mejor los resultados obtenidos en una muestra concreta. De hecho, el conocimiento de estas probabilidades permite utilizar la **distribución multinomial** para conocer la probabilidad concreta asociada a cada posible resultado muestral.

La distribución multinomial

La distribución binomial recién estudiada no es más que un caso especial del caso general: la distribución *multinomial*. La distribución binomial trata con ensayos que únicamente pueden tomar dos resultados (éxito, fracaso); la distribución multinomial permite trabajar con ensayos que pueden tomar cualquier número de resultados.

Supongamos que se realizan n ensayos, cada uno de ellos con I posibles resultados mutuamente exclusivos (por ejemplo, seleccionar una muestra de sujetos y clasificarlos como fumadores, exfumadores o no fumadores; o según su nivel de estudios; etc.). Llamando n_1, n_2, ..., n_I a los I posibles resultados y π_1, π_2, ..., π_I a las probabilidades asociadas a cada posible resultado, la distribución multinomial permite calcular la probabilidad de encontrar una combinación n_i concreta de resultados mediante

$$P(n_i) = \frac{n!}{n_1!\, n_2! \ldots n_I!}\, \pi_1^{n_1}\, \pi_2^{n_2} \ldots \pi_I^{n_I} \qquad [3.10]$$

(siempre que las probabilidades π_i permanezcan constantes en cada ensayo). Para indicar que la variable n_i se distribuye según el modelo multinomial utilizamos la expresión

$$n_i \sim M(n, \pi_i)$$

Aplicando [3.10] a los datos de la Tabla 3.6, donde $n = 50$, $n_1 = 17$, $n_2 = 5$ y $n_3 = 28$, y donde los valores de π_i son, respectivamente, 0,30, 0,05 y 0,65 se obtiene

$$P(n_1 = 17, n_2 = 5, n_3 = 28) = \frac{50!}{17!\, 5!\, 28!}\, (0{,}30)^{17}\, (0{,}05)^{5}\, (0{,}65)^{28} = 0{,}0054$$

Este dato indica que es muy poco probable que los resultados de la Tabla 3.6 se hayan obtenido de una población con las probabilidades propuestas.

Cada posible resultado (cada n_i) puede concebirse como una variable dicotómica pues en cada ensayo puede ocurrir n_i o no ocurrir. Por tanto, cada posible resultado n_i se distribuye binomialmente con valor esperado $E(n_i) = n\pi_i$ y varianza $V(n_i) = \sigma^2_{n_i} = n\pi_i(1-\pi_i)$.

Apéndice 3

Tablas de frecuencias con variables de respuesta múltiple

La expresión *variables de respuesta múltiple* se utiliza para identificar variables en las que los sujetos pueden dar más de una respuesta, es decir, variables en las que un mismo sujeto puede tener valores distintos. En una investigación de tipo social, por ejemplo, podría pedirse a los encuestados: "Señale cuál de los siguientes tipos de transporte público terrestre ha utilizado durante el último mes: autobús, metro, tren, taxi". Obviamente, un mismo sujeto podría marcar más de una respuesta; por esta razón, a la cuestión planteada se le llama variable de respuesta múltiple. Otro ejemplo: en una investigación médica podría pedirse a los pacientes que indicaran cuál de una serie de síntomas han padecido durante la última semana; puesto que los pacientes pueden marcar más de un síntoma, de nuevo la cuestión planteada es una variable de respuesta múltiple.

Al intentar codificar variables de respuesta múltiple surge un problema: los archivos de datos de los programas informáticos (también los archivos de datos del SPSS) sólo permiten utilizar variables con un único código para cada caso. Por esta razón, para poder trabajar con variables de respuesta múltiple es necesario utilizar más de una variable. Esto puede hacerse siguiendo dos estrategias distintas.

Dicotomías múltiples y categorías múltiples

La primera estrategia consiste en crear tantas variables *dicotómicas* como alternativas de respuesta tiene la pregunta. En el ejemplo sobre transporte público habría que crear cuatro variables: *autobús*, *metro*, *tren* y *taxi*. Así, si un encuestado marca la respuesta *autobús*, tendrá en la primera variable un "uno"; si no la marca, un "cero". Cada encuestado tendrá "unos" en las variables que haya marcado y "ceros" en las que no haya marcado. Esta forma de codificar las variables de respuesta múltiple se llama método o estrategia de **dicotomías múltiples**.

La segunda estrategia de codificación de variables de respuesta múltiple consiste en crear tantas variables *categóricas* como respuestas distintas hayan dado los sujetos. Si hay algún sujeto que ha marcado las cuatro respuestas, hay que crear cuatro variables; por ejemplo: *resp1*, *resp2*, *resp3* y *resp4*. Si ningún sujeto ha marcado más de tres respuestas, bastará con crear tres variables; etc. Todas las variables categóricas se codifican ahora con cuatro valores: 1 = «autobús», 2 = «metro», 3 = «tren» y 4 = «taxi». Y a cada sujeto se le asigna, en cada variable, el código correspondiente a su respuesta. Si un sujeto responde, por ejemplo, *autobús* y *taxi*, en la variable *resp1* tendrá un 1 (código correspondiente a *autobús*) y en la variable *resp2* tendrá un 4 (código correspondiente a *taxi*); en el resto de variables tendrá códigos de valor perdido; si un sujeto únicamente responde *metro*, en la variable *resp1* tendrá un 2 (código correspondiente a *metro*) y en el resto de variables códigos de valor perdido; etc. Esta forma de codificar las variables de respuesta múltiple se denomina método o estrategia de **categorías múltiples**.

Aunque toda variable de respuesta múltiple puede codificarse con cualquiera de estas dos estrategias, las características de la propia variable pueden hacer más recomendable una u otra. En la pregunta sobre transportes públicos, puesto que las posibles respuestas son sólo cuatro, el método de dicotomías

múltiples puede resultar tan válido como el de categorías múltiples, y es más rápido y sencillo. Pero cuando el número de posibles respuestas es muy alto (un listado de, por ejemplo, 25 síntomas) y los sujetos sólo marcan unas pocas respuestas, es más apropiado el método de categorías múltiples.

La Figura 3.6 muestra la codificación asignada a las respuestas de una muestra de 20 encuestados a los que se les ha preguntado por el tipo de transporte público terrestre que utilizan. El archivo recoge, en las dos primeras columnas, las variables *id* (número de identificación del caso) y *sexo* (1 = «hombres» y 2 = «mujeres»). A continuación aparecen cuatro variables dicotómicas (*autobús*, *metro*, *tren* y *taxi*) en las que el valor 1 indica que se ha utilizado ese transporte y el valor 0 que no se ha utilizado (método de dicotomías múltiples). Las últimas tres variables ofrecen la misma información que las cuatro variables dicotómicas, pero en formato de categorías múltiples. En este segundo formato, puesto que ningún sujeto ha marcado los cuatro transportes (el que más ha marcado ha marcado tres), sólo es necesario crear tres variables categóricas. El primer sujeto, por ejemplo, ha utilizado el *autobús* y el *tren*, luego en la variable *resp_1* tiene un 1 (código correspondiente a *autobús*) y en la variable *resp_2* tiene un 3 (código correspondiente a *tren*); y como ya no ha marcado ninguna respuesta más, en la variable *resp_3* tiene un 0 (que funciona como código de valor perdido).

Figura 3.6. Datos correspondientes a una muestra de 20 encuestados

	id	sexo	autobús	metro	tren	taxi	resp_1	resp_2	resp_3
1	1	1	1	0	1	0	1	3	0
2	2	1	1	1	0	0	1	2	0
3	3	1	1	1	1	0	1	2	3
4	4	1	1	0	1	0	1	3	0
5	5	1	0	1	1	0	2	3	0
6	6	1	0	0	0	1	4	0	0
7	7	1	1	0	1	0	1	3	0
8	8	1	0	1	1	0	2	3	0
9	9	1	0	1	0	1	2	4	0
10	10	1	1	1	1	0	1	2	3
11	11	2	1	1	0	0	1	2	0
12	12	2	0	1	1	0	2	3	0
13	13	2	0	1	0	0	2	0	0
14	14	2	1	1	1	0	1	2	3
15	15	2	0	1	1	0	2	3	0
16	16	2	1	0	1	0	1	3	0
17	17	2	0	1	0	1	2	4	0
18	18	2	0	1	1	0	2	3	0
19	19	2	1	0	0	1	1	4	0
20	20	2	0	1	1	1	2	3	4

Con un procedimiento SPSS convencional, saber cuántos sujetos utilizan cada medio de transporte requiere describir cada variable de forma individual, por separado. Por ejemplo, con el procedimiento **Estadísticos descriptivos > Frecuencias** del menú **Analizar** se obtienen los resultados que muestra la Tabla 3.7. La tabla informa sobre la frecuencia de uso de cada medio de transporte, pero por separado, es decir, con una tabla distinta para cada medio de transporte. Por el contrario, el procedimiento **Respuestas múltiples** permite ordenar la frecuencia de uso como muestra la Tabla 3.8, donde cada variable dicotómica constituye una categoría de la variable de respuesta múltiple. Veamos cómo hacer esto.

Cómo definir conjuntos de respuestas múltiples

Antes de poder obtener distribuciones de frecuencias como la que muestra la Tabla 3.8 es necesario definir previamente unas nuevas variables llamadas *conjuntos* de respuestas múltiples. El SPSS permite

definir tanto *conjuntos de categorías múltiples* (agrupando variables categóricas) como *conjuntos de dicotomías múltiples* (agrupando variables dicotómicas). Ambos tipos de conjuntos pueden crearse utilizando dos opciones distintas:

1. **Respuesta múltiple > Definir conjuntos de variables** del menú **Analizar**.
2. **Definir conjuntos de respuestas múltiples** del menú **Datos**.

(La segunda opción también ésta disponible en el procedimiento **Tablas** del menú **Analizar**). Al elegir entre estas dos opciones es importante tener en cuenta que los conjuntos que se definan con la segunda opción no podrán utilizarse con los procedimientos **Frecuencias** y **Tablas** de contingencias de la opción **Respuestas múltiples** del menú **Analizar**, y que los conjuntos que se definan con la primera opción tampoco podrán utilizarse con el resto de procedimientos SPSS que permiten trabajar con respuestas múltiples (como, por ejemplo, el procedimiento **Tablas > Tablas personalizadas** del menú **Analizar**). Para obtener tablas de frecuencias y de contingencias es preferible utilizar la primera opción.

Tabla 3.7. Frecuencia de uso del transporte público terrestre: variables tabuladas por separado

Autobús

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	No utilizado	10	50,0	50,0	50,0
	Utilizado	10	50,0	50,0	100,0
	Total	20	100,0	100,0	

Metro

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	No utilizado	6	30,0	30,0	30,0
	Utilizado	14	70,0	70,0	100,0
	Total	20	100,0	100,0	

Tren

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	No utilizado	7	35,0	35,0	35,0
	Utilizado	13	65,0	65,0	100,0
	Total	20	100,0	100,0	

Taxi

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	No utilizado	15	75,0	75,0	75,0
	Utilizado	5	25,0	25,0	100,0
	Total	20	100,0	100,0	

Tabla 3.8. Frecuencia de uso del transporte público terrestre: variables tabuladas juntas

		Respuestas		Porcentaje de casos
		Nº	Porcentaje	
Transporte público	Autobús	10	23,8%	50,0%
	Metro	14	33,3%	70,0%
	Tren	13	31,0%	65,0%
	Taxi	5	11,9%	25,0%
Total		42	100,0%	210,0%

Para definir un conjunto mediante la opción **Respuesta múltiple > Definir conjuntos de variables** del menú **Analizar**, se debe comenzar trasladando a la lista **Variables del conjunto** las variables que se desea incluir en el conjunto (al menos dos).

Las variables individuales que van a formar parte del conjunto pueden estar codificadas como dicotomías o como categorías:

1. **Dicotomías**: debe elegirse esta opción cuando las variables individuales que van a formar parte del conjunto están codificadas como variables dicotómicas (como *autobús*, *metro*, *tren* y *taxi* en la Figura 3.6). En el cuadro de texto **Valor contado** hay que introducir el valor de la variable que debe computarse. Las variables dicotómicas se suelen codificar con "unos" (para la presencia de la característica) y "ceros" (para la ausencia de la característica); por tanto, generalmente, el valor que habrá que introducir en el cuadro de texto **Valor contado** será un «uno». Cada variable que contenga al menos una aparición del valor contado se convierte en una categoría del nuevo conjunto de dicotomías múltiples.
2. **Categorías**. Esta opción debe elegirse cuando las variables individuales que van a formar parte del conjunto están codificadas como variables categóricas (es decir, como *resp_1*, *resp_2* y *resp_3* en el ejemplo de la Figura 3.6). El nuevo conjunto de respuestas múltiples tendrá el mismo rango de valores (categorías) que las variables individuales que lo componen; las cuales, a su vez, deben estar categorizadas de idéntica manera. Para definir ese rango de categorías es necesario introducir, en los cuadros de texto **Rango** y **hasta**, los números enteros que permiten identificar los valores mínimo y máximo entre los que se encuentran los códigos correspondientes a las categorías de las variables individuales.

Cada nuevo conjunto de variables debe tener un nombre único y debe ajustarse a las reglas propias de los nombres de variable del SPSS. También puede asignarse una etiqueta descriptiva del conjunto recién nombrado.

Una vez seleccionadas las variables que formarán parte del conjunto y que se ha asignado nombre al nuevo conjunto, el botón **Añadir** permite hacer efectivas todas estas definiciones incluyendo el nuevo conjunto en la lista **Conjuntos de respuestas múltiples**. Seleccionando un conjunto previamente añadido a esta lista, los botones **Borrar** y **Cambiar** permiten eliminarlo o modificarlo.

Para definir un conjunto de *dicotomías múltiples* utilizando los datos de la Figura 3.6:

- Seleccionar la opción **Respuesta múltiple > Definir conjuntos de variables** del menú **Analizar**.
- Trasladar las variables *autobús*, *metro*, *tren* y *taxi* a la lista **Variables del conjunto**.
- Marcar la opción **Dicotomías** del recuadro **Las variables están codificadas como** e introducir el valor 1 en el cuadro de texto **Valor contado**.
- Para asignar nombre y etiqueta al nuevo conjunto, escribir *trans_d* en el cuadro de texto **Nombre** y *Transporte público (dicotomías)* en el cuadro de texto **Etiqueta** (el sistema añade automáticamente el símbolo "$" delante del nombre asignado al conjunto).
- Pulsar el botón **Añadir** para trasladar (y con ello definir) el conjunto a la lista **Conjuntos de respuestas múltiples**.

Para definir un conjunto de *categorías múltiples*:

- Trasladar las variables *resp_1*, *resp_2* y *resp_3* a la lista **Variables del conjunto**.
- Marcar la opción **Categorías** del recuadro **Las variables están codificadas como** e introducir los valores 1 y 4 en los cuadros de texto **Rango** y **hasta**.
- Para asignar nombre y etiqueta al nuevo conjunto de categorías múltiples, escribir *trans_c* en el cuadro de texto **Nombre** y *Transporte público (categorías)* en el cuadro de texto **Etiqueta** (el sistema añade automáticamente el símbolo "$" delante del nombre asignado al conjunto).
- Pulsar el botón **Añadir** para trasladar (y con ello definir) el conjunto a la lista **Conjuntos de respuestas múltiples**.

A partir de este momento, los conjuntos recién creados (*$trans_d* y *$trans_c*) estarán disponibles en los cuadros de diálogo asociados a la opción **Respuesta múltiple** del menú **Analizar** y podrán utilizarse para obtener tablas de frecuencias y tablas de contingencias.

Cómo obtener tablas de frecuencias

Para obtener tablas de frecuencias con variables de respuesta múltiple hay que seleccionar la opción **Respuestas múltiples > Frecuencias** del menú **Analizar** para acceder al cuadro de diálogo **Frecuencias de respuestas múltiples**. La lista **Conjuntos de respuestas múltiples** ofrece un listado de todos los conjuntos previamente definidos en el cuadro de diálogo **Definir conjuntos de respuestas múltiples** (ver apartado anterior). En este listado no aparecen los conjuntos definidos con otras opciones. Para obtener distribuciones de frecuencias basta con trasladar a la lista **Tablas para** el conjunto o conjuntos que se desea describir.

La opción **Excluir casos según lista dentro de las dicotomías** excluye del análisis los casos con valor perdido en cualquiera de los conjuntos dicotómicos seleccionados; con esta opción desactivada se eliminan del análisis de cada conjunto únicamente los casos con valor perdido en ese conjunto; un caso se considera un valor perdido cuando no puntúa (valor contado) en ninguna de las variables dicotómicas del conjunto. La opción **Excluir casos según lista dentro de las categorías** excluye del análisis los casos con valor perdido en cualquiera de los conjuntos categóricos seleccionados; con esta opción desactivada se eliminan del análisis de cada conjunto únicamente los casos con valor perdido en ese conjunto. Un caso se considera un valor perdido cuando no contiene valores dentro del rango definido en ninguna de las variables categóricas del conjunto.

Para obtener tablas de frecuencias basadas en los dos conjuntos de respuestas múltiples definidos en el apartado anterior:

- Seleccionar la opción **Respuestas múltiples > Frecuencias** del menú **Analizar** para acceder al cuadro de diálogo **Frecuencias de respuestas múltiples**.
- Seleccionar uno de los dos conjuntos previamente definidos, *$trans_d* o *$trans_c* (ver apartado anterior) y trasladarlo a la lista **Tablas para** (los resultados que se obtienen con ambos conjuntos son idénticos).

Aceptando estas selecciones, el *Visor* ofrece los resultados que muestra la Tabla 3.9. Estos resultados corresponden al conjunto de dicotomías múltiples *$trans_d*, el cual se basa en las variables *autobús*, *metro*, *tren* y *taxi*.

El *número de respuestas* indica cuántos sujetos han elegido cada tipo de transporte público. El porcentaje de respuestas indica el porcentaje que cada número de respuestas (10, 14, 13 y 5) representa sobre el total de respuestas (42). El porcentaje de casos se refiere al porcentaje que cada número de respuestas (10, 14, 13, 5) representa sobre el número total de casos (20). El llamativo porcentaje de la última fila (210 %) indica el porcentaje que representa el número total de respuestas (42) sobre el número total de casos.

Tabla 3.9. Frecuencia de uso del transporte público terrestre (dicotomías múltiples)

		Respuestas		Porcentaje de casos
		Nº	Porcentaje	
Transporte público (dicotomías)[a]	Autobús	10	23,8%	50,0%
	Metro	14	33,3%	70,0%
	Tren	13	31,0%	65,0%
	Taxi	5	11,9%	25,0%
Total		42	100,0%	210,0%

a. Conjunto de dicotomias múltiples tabulado en el valor 1.

Ejercicios

3.1. En la siguiente tabla de frecuencias de la variable *X* faltan los nombres de las columnas. A partir de las características de los datos, asignar a cada columna el símbolo que le corresponde y su significado.

X	*a*	*b*	*c*	*d*	*e*	*f*
1	5	0,10	5	10	0,10	10
2	30	0,50	25	60	0,60	50
3	45	0,30	15	90	0,90	30
4	50	0,10	5	100	1,00	10

3.2. Siguiendo con la tabla de frecuencias del ejercicio anterior:

a. Construir el diagrama de barras correspondiente a las frecuencias no acumuladas.
b. ¿Cuál es la moda de la distribución?
c. Calcular el índice de variación cualitativa.

3.3. A continuación se ofrece la tabla de frecuencias (incompleta) obtenida al encuestar a una muestra de 200 personas sobre su actitud hacia la eutanasia (*X*). Completar los valores que faltan (entre paréntesis).

Actitud	n_i	P_i	$\%_i$	na_i	Pa_i	$\%a_i$
Muy en contra	20	( )	( )	( )	( )	( )
En contra	50	( )	( )	( )	( )	( )
Indiferente	30	( )	( )	( )	( )	( )
A favor	80	( )	( )	( )	( )	( )
Muy a favor	20	( )	( )	( )	( )	( )

3.4. A continuación se ofrecen tres tablas de frecuencias (incompletas) obtenidas en tres estudios distintos. Completar los valores que faltan (entre paréntesis) e indicar si en alguna tabla hay algún dato digno de ser destacado.

X	n_i	na_i
1	10	( )
2	15	( )
3	( )	59
4	( )	60
Total	( )	

X	P_i	Pa_i
1	( )	0,15
2	( )	0,35
3	( )	0,85
4	( )	( )
Total	( )	

X	n_i	P_i	Pa_i
1	( )	( )	0,20
2	( )	0,30	( )
3	( )	( )	0,50
4	( )	( )	( )
Total	100	( )	

3.5. Completar la siguiente tabla (valores entre paréntesis) teniendo en cuenta que se ha obtenido al clasificar una muestra de 80 sujetos.

X	n_i	P_i	$\%_i$	na_i	Pa_i	$\%a_i$
1	8	( )	( )	( )	( )	( )
2	( )	0,15	( )	( )	( )	( )
3	( )	( )	( )	( )	0,60	( )
4	( )	( )	25	( )	( )	( )
5	( )	( )	( )	( )	( )	( )

3.6. Un test está formado por 10 preguntas, cada una de las cuales consta de 2 alternativas de respuesta de las que sólo una es correcta. Teniendo en cuenta que la respuesta que se da a cada pregunta es independiente de la respuesta que se da a las demás:

a. Si un sujeto responde al azar a las 10 preguntas del test, ¿qué valores puede tomar la variable n_1 = «número de aciertos» y qué probabilidad tiene asociada cada uno de ellos?

b. ¿Cuál es el valor esperado de n_1? ¿Y la varianza?

c. ¿Cuál es la probabilidad de acertar, por azar, más de 7 preguntas?

d. Si una muestra aleatoria de 200 sujetos responde al azar a las 10 preguntas, ¿cuántos sujetos cabe esperar que acierten más de 7 preguntas?

3.7. Con un tratamiento convencional de la adicción al alcohol se viene obteniendo un 40% de resultados positivos. Un nuevo tratamiento en fase experimental ha conseguido 14 recuperaciones al aplicarlo a 20 pacientes con adicción al alcohol. ¿Qué sugiere este resultado?

3.8. Se sabe que, en una determinada población, la prevalencia de una enfermedad concreta es del 30%. Se cuenta con una prueba diagnóstica con una *sensibilidad* (ofrece un diagnóstico positivo cuando la persona padece la enfermedad) del 90% y una *especificidad* (ofrece un diagnóstico negativo cuando la persona no padece la enfermedad) del 80%.

a. ¿Cuál es la probabilidad de que la prueba ofrezca un diagnóstico incorrecto? (ver ejercicio 2.13).

b. Si se aplica la prueba a una muestra aleatoria de 30 sujetos, ¿cuántos diagnósticos incorrectos cabe esperar encontrar?

c. En esa muestra aleatoria de 30 sujetos, ¿cuál es la probabilidad de que no haya más de 2 diagnósticos incorrectos?

3.9. Un examen de la asignatura *Análisis de Datos* consta de 20 preguntas con 5 alternativas de respuesta de las que sólo una es correcta. Para seleccionar un punto de corte con el que aprobar a los estudiantes se ha utilizado un criterio basado en el número de aciertos por azar. En concreto, se tiene intención de aprobar a los estudiantes que tengan más aciertos que los que cabe esperar simplemente respondiendo al azar. Para ello se ha decidido colocar el punto de corte allí donde obtener por azar un número de aciertos igual o mayor que ese punto tenga una probabilidad menor que 0,05. ¿Dónde habrá que colocar el punto de corte?

3.10. Un test consta de k preguntas dicotómicas (verdadero, falso) e independientes entre sí. Las preguntas acertadas se puntúan con un 1 y las falladas con un 0. La puntuación total en el test

se obtiene sumando las puntuaciones de las *k* preguntas. Señalar la(s) alternativa(s) correcta(s) sabiendo que la varianza de una suma es la suma de las varianzas:

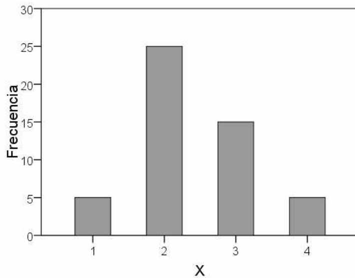
a. La varianza del test vale 0.
b. La varianza del test vale 0,25.
c. La varianza del test vale $k\,(0{,}25)$.
d. La varianza máxima del test vale $k\,(0{,}25)$.
e. La varianza del test depende de la relación que exista entre las preguntas.

3.11. Al encuestar a una muestra aleatoria de 20 personas, cuál es la probabilidad de que más de la mitad se manifieste en contra de la eutanasia sabiendo que, en la población de donde procede esa muestra, el 40% de las personas está en contra de la eutanasia.

Soluciones

3.1. $a = na_i$ (frecuencia absoluta acumulada).
$b = P_i$ (frecuencia relativa).
$c = n_i$ (frecuencia absoluta).
$d = \%a_i$ (frecuencia porcentual acumulada).
$e = P_i$ (frecuencia relativa acumulada).
$f = \%_i$ (frecuencia porcentual).

3.2. *a*. Diagrama de barras:



b. $X = 2$ (categoría con mayor frecuencia no acumulada).

c. $$ICV = \frac{\sum n_i n_{i'}}{I\,(I-1)\,(n/I)^2/2} = \frac{5(25)+5(15)+5(5)+25(15)+25(5)+15(5)}{4\,(4-1)(50/4)^2/2} = \frac{800}{937{,}5} = 0{,}85.$$

3.3.

Actitud	n_i	P_i	$\%_i$	na_i	Pa_i	$\%a_i$
Muy en contra	20	0,10	10	20	0,10	10
En contra	50	0,25	25	70	0,35	35
Indiferente	30	0,15	15	100	0,50	50
A favor	80	0,40	40	180	0,90	90
Muy a favor	20	0,10	10	200	1,00	100

3.4. 1ª tabla: $n = na_4 = 60$; $na_1 = 10$; $na_2 = 10 + 15 = 25$; $n_4 = 60 - 59 = 1$; $n_3 = 59 - 25 = 34$.
2ª tabla: $Pa_4 = 1{,}00$; $P_1 = 0{,}15$; $P_2 = 0{,}35 - 0{,}15 = 0{,}20$; $P_3 = 0{,}85 - 0{,}35 = 0{,}50$; $P_3 = 0{,}85 - 0{,}35 = 0{,}50$.
3ª tabla: $Pa_2 = 0{,}20 + 0{,}30 = 0{,}50$; $Pa_4 = 1{,}00$; $P_1 = 0{,}20$; $P_3 = 0{,}50 - 0{,}50 = 0$; $P_4 = 1 - 0{,}50 = 0{,}50$;
$n_1 = 0{,}20(100) = 20$; $n_2 = 0{,}30(100) = 30$; $n_3 = 0(100) = 0$; $n_4 = 0{,}50(100) = 50$.

En la 1ª tabla hay una categoría cuya frecuencia es muy pequeña ($n_4 = 1$).

En la 3ª tabla hay una categoría cuya frecuencia vale cero (n_3).

3.5. $n = 80$; $P_1 = 8/80 = 0{,}10$; $\%_1 = 0{,}10(100) = 10$; $na_1 = 8$; $Pa_1 = 0{,}10$; $\%a_1 = 10$;
$n_2 = 80(0{,}15) = 12$; $\%_2 = 0{,}15(100) = 15$; $na_2 = 8 + 12 = 20$; $Pa_2 = 0{,}10 + 0{,}15 = 0{,}25$; $\%a_2 = 10 + 15 = 25$;
$P_3 = 0{,}60 - 0{,}25 = 0{,}35$; $n_3 = 0{,}35(80) = 28$; $\%_3 = 0{,}35(100) = 35$; $na_3 = 20 + 28 = 48$; $\%a_3 = 25 + 35 = 60$;
$P_4 = 25/100 = 0{,}25$; $n_4 = 0{,}25(80) = 20$; $na_4 = 48 + 20 = 68$; $Pa_4 = 0{,}60 + 0{,}25 = 0{,}85$; $\%a_4 = 60 + 25 = 85$;
$n_5 = 80 - 68 = 12$; $P_5 = 12/80 = 0{,}15$; $\%_5 = 0{,}15(100) = 15$; $na_5 = 68 + 12 = 80$; $Pa_5 = 1{,}00$; $\%a_5 = 100$.

3.6. *a*. La variable n_1 puede tomar valores entre 0 y 10 (es decir, 0 aciertos, 1 acierto, 2 aciertos, ..., 10 aciertos). Y dado que se trata de la suma de 10 ensayos independientes de una variable dicotómica (1 = «acierto», 0 = «error»), es decir, de la suma de 10 ensayos de Bernoulli independientes entre sí, su distribución es la binomial, por lo que puede utilizarse la tabla B del Apéndice final para conocer la probabilidad asociada a cada valor de n_1. La siguiente tabla ofrece esas probabilidades acumuladas (F) e individuales (f):

n_1	0	1	2	3	4	5	6	7	8	9	10
$F(n_1)$	0,001	0,011	0,055	0,172	0,377	0,623	0,828	0,945	0,989	0,999	1,000
$f(n_1)$	0,001	0,010	0,044	0,117	0,205	0,246	0,205	0,117	0,044	0,010	0,001

b. Ecuación [3.7]: $E(n_1) = n\pi_1 = 10(0{,}50) = 5$.
Ecuación [3.9]: $V(n_1) = n\pi_1(1-\pi_1) = 10(0{,}50)(0{,}50) = 2{,}5$.

c. De acuerdo con los datos de la distribución de n_1 obtenida en el apartado *a*:
$P(n_1 > 7) = 1 - F(7) = 1 - 0{,}945 = 0{,}055$.

d. $100\,P(n_1 > 7) = 100\,(0{,}055) = 11$ sujetos.

3.7. Vamos a calcular la probabilidad de obtener con el tratamiento convencional el resultado obtenido con el experimental. Si esa probabilidad es muy baja, significará que el resultado obtenido con el tratamiento experimental es muy difícil de obtener con el convencional.

Tenemos $n = 20$ ensayos de una variable dicotómica (1 = «recuperación», 0 = «no recuperación») con probabilidad de recuperación $\pi_1 = 0{,}40$ en cada ensayo. Tenemos, por tanto, una variable que se distribuye binomialmente. Consultando la tabla de la distribución binomial (Tabla B del Apéndice final) obtenemos: $P(n_1 > 13) = 1 - F(13) = 1 - 0{,}994 = 0{,}006$. Por tanto, es muy poco probable (0,006) que el resultado obtenido con el tratamiento experimental pueda obtenerse con el convencional.

3.8. *a*. El diagnóstico será incorrecto tanto si la prueba dice que una persona que está enferma (E) no lo está (−), como si dice que una persona que no está enferma (nE) lo está (+):

- $P(E \cap -) = P(E)\,P(- \mid E) = 0{,}30 \times 0{,}10 = 0{,}03$.
- $P(nE \cap +) = P(nE)\,P(+ \mid nE) = 0{,}70 \times 0{,}20 = 0{,}14$.

Por tanto, la probabilidad de diagnóstico equivocado será: $0{,}03 + 0{,}14 = 0{,}17$.

b. Tenemos $n = 30$ ensayos de una variable dicotómica (1 = «diagnóstico incorrecto», 0 = «diagnóstico correcto») con probabilidad de diagnóstico incorrecto $\pi_1 = 0{,}17$ en cada ensayo. Lo que cabe

esperar que ocurra es el valor esperado de la variable n_1 = «número de diagnósticos incorrectos». Es decir, $E(n_1) = n\pi_1 = 30(0,17) = 5,1$ (ecuación [3.7]). Por tanto, en 30 diagnósticos, lo más probable es que la prueba se equivoque en 5.

c. La probabilidad de que no haya más de 2 diagnósticos incorrectos es la probabilidad de que haya dos, uno o ninguno; es decir:

$P(n_1 < 3) \ = \ P(n_1 = 0) + P(n_1 = 1) + P(n_1 = 2)$

Aplicando la ecuación [3.4] con $n = 30$ y $\pi_1 = 0,17$ se obtiene:

- $P(n_1 = 0) \ = \ [30! \,/\, (0!\ 30!)]\ 0,17^0\ (0,83)^{30} \ = \ 1(0,003735) \ = \ 0,003735.$
- $P(n_1 = 1) \ = \ [30! \,/\, (1!\ 29!)]\ 0,17^1\ (0,83)^{29} \ = 30\,(0,000765) \ = \ 0,022953.$
- $P(n_1 = 2) \ = \ [30! \,/\, (2!\ 28!)]\ 0,17^2\ (0,83)^{28} \ = 435\,(0,000157) \ = \ 0,068167.$

Por tanto: $P(n_1 < 3) \ = \ 0,003735 + 0,022953 + 0,068167 \ = \ 0,095.$

3.9. Tenemos una variable dicotómica (1 = «acierto», 0 = «error») de la que se realizan $n = 20$ ensayos. La probabilidad de acierto en cada ensayo vale $\pi_1 = 1/5 = 0,20$. En la tabla de la distribución binomial (Tabla B del Apéndice final) puede comprobarse:

$P(n_1 = 7) \ = \ 1 - 0,913 \ = \ 0,087.$
$P(n_1 = 8) \ = \ 1 - 0,968 \ = \ 0,032.$

La probabilidad de obtener 7 o más aciertos es mayor que 0,05 (por tanto, 7 aciertos no sirve como punto de corte). Sin embargo, la probabilidad de obtener 8 o más aciertos es menor que 0,05. La decisión será, por tanto, aprobar a los estudiantes que obtengan 8 aciertos o más.

3.10. La suma de las puntuaciones del test no es más que el número de aciertos (n_1). Y la varianza del número de aciertos (ver ecuación [3.9]) viene dada por $V(n_1) = n\pi_1(1 - \pi_1)$, es decir, $k\pi_1(1 - \pi_1)$. Ahora bien, el producto entre dos números que suman 1 es máximo cuando esos números son iguales. Por tanto la varianza de n_1 será máxima cuando $\pi_1 = 1 - \pi_1 = 0,50$. Luego la alternativa *d* es correcta: el valor máximo de la varianza del test (n_1) es $k\,(0,25)$. La varianza del test sólo valdrá cero cuando todas las respuestas sean aciertos o todas sean fallos; y no sabemos cuáles son las respuestas (esto descarta la alternativa *a*). Y las alternativas *b* y *c* se refieren a valores concretos que no están justificados, pues no se dispone de ninguna información para saber cuál es el valor concreto de la varianza.

3.11. Tenemos una variable dicotómica (1 = «en contra», 0 = «a favor») de la que se han realizado $n = 20$ ensayos. Siendo n_1 = «número de personas que se manifiestan en contra de la eutanasia» y π_1 = «proporción de personas en contra de la eutanasia», las probabilidades asociadas a la variable n_1 pueden obtenerse en la tabla de la distribución binomial (Tabla B del Apéndice final):

$P(n_1 > 10) \ = \ 1 - F(10) \ = \ 1 - 0,872 \ = \ 0,128.$

4

Análisis descriptivo de variables cuantitativas

En el capítulo anterior hemos comenzado señalando que, cualquiera que sea el tipo de análisis que finalmente se tenga intención de llevar a cabo para poder cubrir los objetivos de un estudio, lo primero que conviene hacer con un conjunto de datos es formarse una idea lo más completa posible acerca de sus características. Esto sigue siendo válido cuando se trabaja con datos cuantitativos. Por tanto, antes que nada, lo primero que conviene hacer es familiarizarse con las herramientas que permiten formarse esa primera idea sobre las características de los datos; es decir, antes que nada, conviene conocer las herramientas diseñadas para explorar los datos y obtener información descriptiva.

Acabamos de ver que un resumen tan simple como una tabla o distribución de frecuencias contiene gran parte de la información necesaria para caracterizar una variable categórica; la propia naturaleza de la variable hace innecesario realizar cálculos más allá de un recuento o un porcentaje. Con variables cuantitativas, sin embargo, una tabla de frecuencias no es más que un listado que no resume nada o casi nada y, consecuentemente, una herramienta muy poco informativa: demasiados valores distintos y, si la medida utilizada es lo bastante precisa, muy pocas repeticiones. Organizar y resumir (es decir, describir) los valores de una variable cuantitativa requiere utilizar herramientas algo más sofisticadas. Pero sigue siendo necesario prestar atención a tres propiedades básicas: *centro*, *dispersión* y *forma de la distribución*.

Los estadísticos diseñados para identificar el *centro* de una variable suelen agruparse bajo la denominación de medidas de **tendencia central**. Los estadísticos que permiten cuantificar el grado de *dispersión* (alejamiento) de las puntuaciones respecto de su centro suelen agruparse bajo la denominación de medidas de **dispersión**, aunque también pueden encontrarse con la denominación de medidas de *variación* o de *variabilidad*. Y los estadísticos que sirven para describir la *forma de la distribución* suelen prestar atención a dos características: **asimetría** y **curtosis**.

En este capítulo se presentan los estadísticos habitualmente utilizados para describir el centro, la dispersión y la forma de la distribución de las variables cuantitativas, más otro tipo de estadísticos cuyo objetivo principal no es el de resumir la información, sino el de ubicar a los sujetos individualmente considerados en la posición relativa que ocupan respecto del resto de sujetos de su grupo: los ***cuantiles***.

Cuantiles

Aunque la mayoría de las herramientas descriptivas que estudiaremos en este y otros capítulos están diseñadas para organizar y resumir conjuntos de datos, existe un tipo particular de herramientas que, además de ayudar a organizar y resumir conjuntos de datos, también sirven para describir datos individuales.

¿Por qué decimos que un cociente intelectual de 140 puntos es alto, o que un jugador de baloncesto que mide 175 cm es bajo? La respuesta es clara: porque muy pocas personas alcanzan un conciente intelectual de 140 puntos, y porque la mayoría de los jugadores de baloncesto miden más de 175 cm. Es decir, afirmamos que una puntuación es alta o baja porque la comparamos con el resto de puntuaciones del conjunto al que pertenece.

Comparar puntuaciones entre sí permite ordenarlas por su tamaño y, una vez ordenadas, ya es posible conocer la posición relativa que cada una ocupa dentro del conjunto. De hecho, existen varios estadísticos que cumplen con esta función de posicionamiento relativo tomando como referencia la ordenación ascendente de las puntuaciones (ránking de menor a mayor). Estos estadísticos reciben el nombre genérico de *cuantiles*.

Los **cuantiles** son cada uno de los J valores que dividen la distribución en $J+1$ partes iguales. Algunos cuantiles tienen nombre específico. Así, la **mediana** es *un* valor que divide la distribución en *dos* partes iguales (la mitad de los casos en cada parte; volveremos enseguida sobre ella). Los **cuartiles** son *tres* valores (Q_1, Q_2, Q_3) que dividen la distribución en *cuatro* partes iguales (el 25% de los casos en cada parte). Los **deciles** son *nueve* valores (D_1, D_2, ..., D_9) que dividen la distribución en *diez* partes iguales (el 10% de los casos en cada parte). Los **centiles** son *noventa y nueve* valores (C_1, C_2, ..., C_{99}) que dividen la distribución en *cien* partes iguales (el 1% de los casos en cada parte).

El concepto de cuantil está estrechamente asociado al de *frecuencia porcentual acumulada*. La mediana, por ejemplo, es el valor que acumula el 50% de los casos; el primer cuartil (Q_1) es el valor que acumula el 25% de los casos; el octavo decil (D_8) es el valor que acumula el 80% de los casos; el quinto centil (C_5) es el valor que acumula el 5% de los casos; etc. De ahí que, al trabajar con cuantiles, lo habitual sea utilizar **percentiles** (P_k), que son, precisamente, valores que acumulan un determinado porcentaje de casos: el percentil P_k es el valor que acumula el k% de los casos.

Conviene advertir que los cuantiles no son porcentajes, sino *valores* de la variable. Si se está trabajando con la variable edad, un cuantil es una edad. Así, si el percentil 25 (P_{25}) de las edades de un grupo de sujetos vale 32,75 años ($P_{25} = 32{,}75$), entonces el 25% de los sujetos de ese grupo tiene menos de 32,75 años (o el 75% de los sujetos más de 32,75 años).

Aunque algunos cuantiles tienen nombre concreto (mediana, cuartiles, deciles, centiles), lo cierto es que todos ellos pueden ser concebidos como percentiles: la mediana es el percentil 50; los cuartiles son los percentiles 25, 50 y 75; los deciles son los percentiles 10, 20, ..., 90; los centiles son los percentiles 1, 2, 3, ..., 99. Por tanto, para conocer el valor de un cuantil cualquiera, basta con calcular el correspondiente percentil. Ahora bien, no existe una única manera de calcular percentiles: la literatura estadística recoge al menos cinco métodos distintos de cálculo. No obstante, el más extendido (el que utilizan programas informáticos como el SPSS) se ajusta a una sencilla regla. Se comienza ordenando los casos de forma ascendente por su valor en Y, y calculando $i = k(n+1)/100$; a continuación:

– Si i es un número entero, entonces $P_k = Y_i$
– Si i es un número decimal, entonces $P_k = (1-d)\,Y_i + (d)\,Y_{i+1}$ [4.1]

donde k es el porcentaje de casos acumulados que se busca; Y_i es el valor de la variable que ocupa la posición correspondiente a la parte entera de i; y d es la parte decimal de i.

Veamos cómo se calcula el percentil 25 (P_{25}) o primer cuartil (Q_1) con las edades de la Tabla 4.1. Para calcular la edad que corresponde al P_{25} en el grupo de hombres, comenzamos calculando la posición que corresponde a ese percentil en ese grupo:

$$i = 25(11+1)/100 = 3$$

Como el valor de i es un número entero, $P_{25} = Y_i$. Es decir, el percentil 25 es el valor (la edad) que ocupa la 3ª posición; por tanto, en el grupo de hombres, $P_{25} = 20$ años.

Para obtener el P_{25} en el grupo de mujeres comenzamos calculando la posición que corresponde al P_{25} en ese grupo:

$$i = 25(14+1)/100 = 3{,}75$$

Como el valor de i no es un número entero, es necesario interpolar mediante [4.1] tomando $i = 3$ y $d = 0{,}75$:

$$P_{25} = (1-0{,}75)\,Y_3 + (0{,}75)\,Y_4 = (0{,}25)21 + (0{,}75)22 = 21{,}75$$

Por tanto, en el grupo de mujeres, $P_{25} = 21{,}75$ años. Este método de cálculo se conoce como *haverage* y, según hemos señalado ya, no es el único disponible[1]. No obstante, con todos ellos se obtienen resultados muy parecidos.

Una de las principales aplicaciones de los cuantiles tiene que ver con la elaboración de tablas de clasificación o *baremos*. Por ejemplo, si los resultados de una determinada prueba diagnóstica (o las alturas de un grupo de sujetos, o sus pesos, etc.) se transforman en percentiles, a partir de ese momento cada resultado individual podrá ser ubicado (clasificado) en la posición relativa que le corresponde en el conjunto de los resultados de la prueba (o en el conjunto de las alturas del grupo, o en el conjunto de los pesos, etc.).

Los cuantiles también sirven para comparar posiciones relativas entre grupos, lo cual es especialmente útil cuando se comparan distribuciones que no son idénticas o variables con distinta métrica. Así, si un estudiante obtiene un 6 en matemáticas en un grupo donde pocos sujetos superan esa nota y otro estudiante obtiene también un 6 en matemáticas en un grupo donde la mayoría de los sujetos superan esa nota, al primer estudiante le corresponderá un percentil más alto que al segundo, lo que estará indicando que el primer estudiante ocupa en su grupo una posición más alta que el segundo en el suyo (volveremos sobre este asunto en el próximo capítulo al hablar de las puntuaciones típicas).

Tabla 4.1. Edades de dos grupos de sujetos (puntuaciones ordenadas para el cálculo de los cuantiles)

Grupos	*Edades*														n_j
Hombres	18	19	20	22	24	26	29	31	35	45	61				11
Mujeres	19	20	21	22	23	24	25	27	28	29	30	31	32	35	14
Posición (i)	1ª	2ª	3ª	4ª	5ª	6ª	7ª	8ª	9ª	10ª	11ª	12ª	13ª	14ª	

[1] Además del método *haverage*, el SPSS incluye (disponibles mediante sintaxis) otros cuatro métodos de cálculo de cuantiles: *waverage*, *round*, *empirical* y *aempirical* (ver, en el apéndice de este mismo capítulo, el apartado *Métodos para el cálculo de cuantiles*).

Tendencia central

Una buena manera de identificar el *centro* de una variable consiste en elegir el valor que mejor representa al resto de valores. En el capítulo anterior hemos elegido como mejor representante de una variable categórica el valor que más se repite (la *moda*). Pero en una variable cuantitativa se dan muy pocas repeticiones (sobre todo si se mide con la suficiente precisión) y el valor que más se repite no tiene por qué estar en el centro; consecuentemente, la moda, además de no aportar información útil, puede resultar engañosa.

El centro de una variable cuantitativa hay que intentar encontrarlo de otra manera. Y lo cierto es que existen diferentes formas de hacerlo; todo depende de qué aspectos de la variable se consideren relevantes: puede prestarse atención a todos los valores de la variable o sólo a parte de ellos; si se decide no prestar atención a todos los valores, la decisión puede atender a distintos criterios; pueden ponderarse todos los valores por igual o pueden asignarse ponderaciones distintas a valores distintos, etc. No existe un estadístico perfecto para describir el centro de todas las variables en el sentido de que no existe ningún estadístico capaz de captar toda la complejidad de una variable cuantitativa; cada estadístico se centra en un aspecto de la variable y ese detalle le confiere sus fortalezas y sus debilidades. Según tendremos ocasión de comprobar, una de las principales tareas que tendremos que acometer será la de aprender a elegir el estadístico apropiado.

Media aritmética

La *media aritmética* (o, simplemente, *media*) es, sin duda, el estadístico de tendencia central más utilizado. Se define como la *suma de todas las puntuaciones dividida por el número de puntuaciones* y se representa con el nombre de la variable coronado con un guión. Así, la media de la variable Y se representa y define mediante

$$\bar{Y} = \frac{\sum Y_i}{n} \qquad [4.2]$$

Consideremos de nuevo los dos grupos de edades de la Tabla 4.1. Llamando Y_h a la edad de los hombres e Y_m a la de las mujeres, la edad media de cada grupo se obtiene de la siguiente manera:

$$\bar{Y}_h = (18+19+20+\cdots+61)/11 = 30{,}00.$$
$$\bar{Y}_m = (19+20+21+\cdots+35)/14 = 26{,}14.$$

Estos valores indican dónde se encuentra el centro de la distribución de las edades de cada grupo: en el de hombres se encuentra en torno a 30 años; en el de mujeres, en torno a 26 años (debe tenerse en cuenta que el valor de la media no tiene por qué coincidir exactamente con ninguno de los valores concretos que toma la variable).

Como centro de la distribución que es, la media está identificando el valor de la distribución en torno al cual cabe esperar encontrar más valores. Pero, por sí sola, tiene una capacidad descriptiva bastante limitada. Por un lado, el valor de la media no dice nada acerca de lo bien o mal que está representando al resto de valores; para saber algo sobre esto hace falta recabar información adicional relacionada con el grado de dispersión del conjunto de valores (la me-

dia será tanto mejor representante del resto de valores cuanto menos dispersos –más concentrados– se encuentren éstos). Por otro lado, el hecho de que en el cálculo de la media intervengan todos los valores hace de ella un estadístico muy sensible a la presencia de asimetría en la distribución, es decir, a la presencia de valores muy alejados del centro por uno de los dos extremos de la distribución (enseguida veremos algunos estadísticos que intentan obtener el centro de una variable sin tener en cuenta todos sus valores).

Propiedades de la media

A pesar de sus limitaciones descriptivas, la media posee algunas importantes propiedades. Una de ellas, que interesa destacar de forma especial porque tiene relación con operaciones que se realizan con mucha frecuencia al analizar datos, se refiere a las diferencias entre cada puntuación y la media:

$$y = Y - \bar{Y} \qquad [4.3]$$

A estas diferencias se les llama **puntuaciones diferenciales** o **puntuaciones de desviación**. Las representamos con letras minúsculas para distinguirlas de las puntuaciones originales o *puntuaciones directas* (Y). Las puntuaciones diferenciales tienen una importante propiedad: *su suma vale cero*. Esto significa que las desviaciones negativas (las asociadas a los valores menores que la media) se anulan con las positivas (las asociadas a los valores mayores que la media); es decir, la suma de las puntuaciones diferenciales negativas (en valor absoluto) es idéntica a la suma de las puntuaciones diferenciales positivas. Esta propiedad de la media permite concebirla como el *centro de gravedad* de la variable: las desviaciones de la media *pesan* exactamente lo mismo a cada lado de la media.

Volvamos a los datos de la Tabla 4.1, referidos a las edades de dos grupos. La edad media del grupo de hombres (la hemos calculado unos párrafos más arriba) vale 30. Aplicando la ecuación [4.3] a esas 11 edades se obtiene

$$\begin{array}{lll} y_1 = 18 - 30 = -12 & y_5 = 24 - 30 = -6 & y_9 = 35 - 30 = 5 \\ y_2 = 19 - 30 = -11 & y_6 = 26 - 30 = -4 & y_{10} = 45 - 30 = 15 \\ y_3 = 20 - 30 = -10 & y_7 = 29 - 30 = -1 & y_{11} = 61 - 30 = 31 \\ y_4 = 22 - 30 = -8 & y_8 = 31 - 30 = 1 & \end{array}$$

$$\sum_i y_i = -12 + (-11) + (-10) + (-8) + (-6) + (-4) + (-1) + 1 + 5 + 15 + 31 = 0.$$

Dedicar tiempo a este sencillo cálculo no tiene otra finalidad que la de poder constatar y grabar bien en la memoria que, efectivamente, unas puntuaciones diferenciales son positivas y otras negativas, y que se anulan al sumarlas. Tendremos ocasión de volver repetidamente sobre estas puntuaciones.

Otra interesante propiedad de la media tiene que ver con los cambios que experimenta cuando se transforman las puntuaciones originales. En concreto, *si a las puntuaciones de una variable se les multiplica y/o se les suma una constante, la media de la variable también queda multiplicada y/o sumada por esa misma constante*. Veamos con algo más de detalle en qué consiste esta propiedad. Siendo a, b y c tres constantes cualesquiera:

$$\text{Si } X = a + Y, \quad \text{entonces } \bar{X} = a + \bar{Y} \qquad [4.4.a]$$

Es decir, si a las edades de los 11 hombres de la Tabla 4.1 (cuya media vale 30) se les suman 10 años, la media de las nuevas edades pasa a valer $30 + 10 = 40$.

$$\text{Si } X = bY, \quad \text{entonces } \bar{X} = b\bar{Y} \qquad [4.4.b]$$

Es decir, si las edades de los 11 hombres de la Tabla 4.1 (cuya media vale 30) se multiplican por 5, la media de las nuevas puntuaciones pasa a valer $30 \times 5 = 150$. La combinación de las dos reglas anteriores tiene como resultado una nueva regla:

$$\text{Si } X = a + bY, \quad \text{entonces } \bar{X} = a + b\bar{Y} \qquad [4.4.c]$$

Es decir, si las edades de los 11 hombres de la Tabla 4.1 (cuya media vale 30) se multiplican por 5 y al producto resultante se le suma 10, la media de las nuevas puntuaciones pasa a valer $10 + 30 \times 5 = 160$. Esta propiedad relativa a los cabios que experimenta la media cuando se transforman las puntuaciones originales se mantiene incluso cuando se combinan dos o más variables:

$$\text{Si } X = aV + bY + \cdots + cZ, \quad \text{entonces } \bar{X} = a\bar{V} + b\bar{Y} + \cdots + c\bar{Z} \qquad [4.4.d]$$

Si las constantes se igualan a 1 se entenderá fácilmente esta regla: la media de una suma de variables es la suma de las medias de las variables. Si, además, las variables sumadas se multiplican por una constante distinta de 1, la media de esas variables quedará multiplicada por la constante y esto quedará igualmente reflejado en la media combinada (debe tenerse en cuenta que esta regla asume que todas las variables están medidas en el mismo conjunto de casos; otra cosa no tendría sentido).

Media ponderada

Cuando se trabaja con varios grupos resulta útil conocer la correspondencia existente entre las *medias de los grupos* y la *media total* (la media de todas las puntuaciones). Si todos los grupos tienen el mismo tamaño, la media total coincide con la media de las medias de los grupos. Si los grupos tienen distinto tamaño, esta correspondencia no se sostiene. En estos casos, la media total es la *media ponderada* de las medias de los grupos. Así, siendo J el número de grupos y $n_1, n_2, ..., n_j, ..., n_J$ el tamaño de los grupos, la media ponderada viene dada por:

$$\bar{Y}_{\text{ponderada}} = \frac{n_1\bar{Y}_1 + n_2\bar{Y}_2 + \cdots + n_J\bar{Y}_J}{n_1 + n_2 + \cdots + n_J} = \frac{\sum n_j\bar{Y}_j}{\sum n_j} \qquad [4.5]$$

(el subíndice j se refiere a cada uno de los grupos: $j = 1, 2, ..., J$). En los datos de la Tabla 4.1 hay 11 hombres con una edad media de 30,00 años y 14 mujeres con una edad media de 26,14 años. Para obtener la edad media de los 25 sujetos a partir de las medias de los grupos es necesario utilizar la media ponderada:

$$\bar{Y}_{\text{ponderada}} = \frac{11(30{,}00) + 14(26{,}14)}{(11+14)} = 27{,}84$$

Este valor coincide con el que se obtiene a partir de las edades individuales de los 25 sujetos y difiere del que se obtiene promediando las medias de los dos grupos sin tener en cuenta su tamaño muestral: $(30{,}00 + 26{,}14)/2 = 28{,}07$. No obstante, si los tamaños son parecidos o las medias difieren poco entre sí, con ambas estrategias se obtendrá un resultado similar.

Mediana

La mediana (Mdn_Y) de una variable (Y) es el centro de la variable en sentido literal: *es el valor que ocupa la posición central cuando los casos están ordenados*. La mediana es, por tanto, el valor que deja por debajo de sí el mismo número de casos que por encima. Con otras palabras, el valor que deja por debajo de sí el 50% de los casos; es decir, el percentil 50.

A pesar de su aparente simplicidad conceptual, los métodos disponibles para calcular la mediana son muy diversos, lo cual está indicando que el cálculo de la mediana no está, ni mucho menos, libre de problemas (particularmente cuando existen valores empatados en torno al centro de la variable). No obstante, el método de cálculo más extendido[2] se ajusta a una regla bastante simple (en todo momento se está asumiendo que las n puntuaciones están ordenadas de menor a mayor: $Y_1 \leq Y_2 \leq \ldots \leq Y_i \leq \ldots \leq Y_n$):

- Si el número de casos es impar, la mediana es el valor que ocupa la posición $i = (n + 1)/2$;
- Si el número de casos es par, la mediana es el punto medio (media aritmética) entre los dos valores que ocupan las posiciones $i_1 = n/2$ e $i_2 = (n/2) + 1$. [4.6]

Para ilustrar el cálculo de la mediana, la Tabla 4.2 ofrece cuatro grupos de puntuaciones ordenadas de menor a mayor. Consideremos los grupos 1 y 3. Como el número de casos del primer grupo es impar ($n_1 = 11$), la mediana es el valor que ocupa la posición $i = (11+1)/2 = 6$ (es decir, la 6ª posición); por tanto, $Mdn_{Y(1)} = 25$. Como el número de casos del tercer grupo es par ($n_3 = 14$), la mediana es la media aritmética de los dos valores que ocupan las posiciones $i_1 = 14/2 = 7$ e $i_2 = (14/2) + 1 = 8$ (posiciones 7ª y 8ª); por tanto, $Mdn_{Y(3)} = (25 + 27)/2 = 26$.

Tabla 4.2. Edades de cuatro grupos de sujetos (puntuaciones ordenadas para el cálculo de la mediana)

Grupos	*Edades*														n_j
1	19	20	22	23	24	25	26	27	28	29	33				11
2	18	19	20	22	24	26	29	31	35	45	61				11
3	19	20	21	22	23	24	25	27	28	29	30	31	32	35	14
4	18	19	20	22	24	26	29	33	35	36	41	43	59	74	14
Posición (i)	1ª	2ª	3ª	4ª	5ª	6ª	7ª	8ª	9ª	10ª	11ª	12ª	13ª	14ª	

Estadísticos resistentes

Se dice que un estadístico es resistente cuando es poco sensible a la presencia de anomalías en los datos. Esta propiedad es particularmente útil cuando se trabaja con distribuciones asimétricas, es decir, con distribuciones que contienen casos muy alejados del centro por uno de los dos extremos de la distribución.

[2] Esta forma de calcular la mediana se conoce como método *haverage* y es el método que utilizan por defecto la mayoría de los programas informáticos. Pero estos programas suelen incluir otros métodos de cálculo. El SPSS, por ejemplo, incluye (mediante sintaxis) otros cuatro métodos. Ver, en el Apéndice 4, el apartado *Métodos para el calculo de cuantiles*.

La media aprovecha las propiedades *cuantitativas* (de intervalo o razón) de los datos. La mediana, sin embargo, sólo aprovecha sus propiedades *ordinales*. Las implicaciones de esta diferencia son importantes. La mediana del primer grupo de edades de la Tabla 4.2 no se altera si el valor 33 se cambia, por ejemplo, por el valor 74. Sin embargo, ese cambio hace que la media del grupo 1 pase de 25,09 a 28,82. Es decir, mientras que el cambio de un sólo dato altera la media en más de tres puntos y medio, la mediana permanece inalterada (el cambio en la media podría ser incluso mayor sin que la mediana experimentara ningún tipo de alteración). La mediana, por tanto, es un estadístico más resistente que la media.

Pero la mediana no es el único estadístico de los llamados *resistentes*. La **media recortada** (o *media truncada*; del inglés *trimmed mean*; ver Hoaglin, Mosteller y Tukey, 1983) se obtiene calculando la media aritmética tras eliminar de los extremos de la distribución un determinado porcentaje de casos. La media recortada al k% es la media aritmética calculada tras eliminar el k% de los casos con valores más pequeños y el k% de los casos con valores más grandes[3] (k es un porcentaje que suele oscilar entre 5 y 25; cuanto mayor es el porcentaje de casos que se recorta, más resistente es el estadístico a la presencia de anomalías en los datos). El objetivo de esta corrección es hacer que el resultado sea menos sensible (más resistente) a la presencia de valores atípicos por uno de los dos extremos de la distribución. Así, mientras la media aritmética de las puntuaciones del cuarto grupo (Tabla 4.2) vale 34,21, la media recortada al 5% vale 32,90 (un valor más cercano a la mediana[4], que en este grupo de puntuaciones vale 31). Obviamente, el valor de la media recortada será tanto más parecido al de la media aritmética cuanto más simétrica sea la distribución.

La **media winsorizada** (Miller, 1986) se basa en una lógica similar a la de la recortada, pero en lugar de eliminar un determinado porcentaje de casos de los extremos, sustituye los valores de esos casos por el valor adyacente a ellos. Así, la media *winsorizada* al k% se obtiene calculando la media aritmética tras sustituir el k% de los valores más pequeños por el valor más pequeño de los restantes y el k% de los valores más grandes por el valor más grande de los restantes. Con esta estrategia se obtiene un estadístico ($\bar{Y}_W$) más resistente que la media aritmética pero algo menos resistente que la media recortada.

La **trimedia** (ver Hoaglin, Mosteller y Tukey, 1983) es un estadístico muy resistente cuyo nombre deriva de su definición: se obtiene *promediando tres valores*: los tres cuartiles[5]. La trimedia también puede encontrarse con el nombre de *BESA* (***b**est **e**asy **s**ystematic **a**verage*) y con el de *mediana recortada*. Se calcula dando al segundo cuartil (la mediana) doble peso:

---

[3] Si $t = kn/100$ es un número entero, la media recortada se obtiene eliminando del análisis los t valores más pequeños y los t valores más grandes. Si t no es un número entero, se separa la parte entera (t_e) de la parte decimal (t_d), se eliminan del análisis los t_e valores más pequeños y los t_e valores más grades, y los dos valores extremos no eliminados se ponderan con $1 - t_d$. Así, para obtener la media recortada al 5% con las puntuaciones del grupo 4 (ver Tabla 4.2) se comienza calculando $t = 5(14)/100 = 0{,}7$. La parte entera de t vale 0; por tanto, no hay que eliminar ningún valor. La parte decimal de t vale 0,7; por tanto, la media recortada se obtiene ponderando los valores extremos 18 y 74 con $1 - 0{,}7 = 0{,}3$:

$$\bar{Y}_{5\%} = \frac{18(0{,}3) + 19 + \cdots + 59 + 74(0{,}3)}{14[100 - 2(5)]/100} = \frac{414{,}6}{12{,}6} = 32{,}90$$

El denominador de la media recortada se obtiene multiplicando n por la proporción de casos que intervienen en el análisis: $n(1 - 0{,}02k)$. En el ejemplo, el denominador se obtiene multiplicando $14 \times 0{,}90$.

[4] En realidad, la mediana puede concebirse como una media recortada al 50%, es decir, como una media que se calcula después de eliminar los casos que quedan por encima y por debajo del valor central.

[5] La formulación original de la *trimedia* no se basa en los cuartiles, sino los *cuartos* o *bisagras* de Tukey (1977). El primer cuarto o bisagra es el valor que ocupa la posición intermedia entre la mediana y el valor más pequeño; el segundo es la mediana; el tercero es el valor que ocupa la posición intermedia entre la mediana y el valor más grande.

$$\bar{Q} = \frac{Q_1 + 2Q_2 + Q_3}{4} \qquad [4.7]$$

Por último, los **estimadores *M*** son un grupo de estadísticos resistentes basados en el método de *máxima verosimilitud* (de ahí la *M* con la que son conocidos). También se les conoce como *estimadores robustos centrales*. Un estimador robusto central o estimador *M* no es más que una media ponderada en la que los pesos asignados a los casos dependen de la distancia de cada caso al centro de la distribución: los casos próximos a la posición central reciben un peso de 1 y los demás valores reciben un peso tanto menor cuanto más alejados se encuentran de la posición central. Existen varios estimadores *M* que difieren entre sí por la forma concreta de asignar pesos a los casos. El SPSS (en el procedimiento **Explorar**) incluye cuatro de estos estimadores: Huber, Tukey, Hampel y Andrews (puede encontrarse una descripción detallada y asequible de todos ellos en Norušis y SPSS Inc., 1993, págs. 192-194; y en Palmer, 1999, págs. 124-162). Andrews, Bickel, Hampel y cols. (1972) recomiendan utilizar el estimador *M* de Andrews y el de Hampel, al tiempo que desaconsejan el uso del de Huber. Al igual que ocurre con el resto de estadísticos resistentes, los estimadores *M* son menos sensibles que la media aritmética a la presencia de valores extremos. Por tanto, cuando las distribuciones son asimétricas, estos estimadores robustos son mejores estadísticos de tendencia central que la media aritmética.

Comparación entre estadísticos de tendencia central

De lo estudiado hasta ahora en este capítulo cabe deducir que uno de los principales objetivos del análisis descriptivo consiste en identificar el centro de una variable. Pero acabamos de ver que este objetivo puede alcanzarse utilizando diferentes estrategias (diferentes estadísticos[6]). El grado de parecido entre estos estadísticos depende, básicamente, de la *forma de la distribución* de la variable: si la distribución es simétrica, todos los estadísticos toman el mismo valor; la diferencia entre ellos va aumentando conforme aumenta el grado de asimetría.

La *media aritmética* utiliza las propiedades cuantitativas de los datos y se basa en una ponderación uniforme de todos ellos. Esto la convierte en un estadístico muy sensible (poco resistente) a la presencia de asimetría en la distribución de los datos. Las medias *recortada* y *winsorizada* intentan corregir la falta de resistencia de la media aritmética modificando el tratamiento que dan a un determinado porcentaje de casos de los extremos de la distribución;

---

[6] Todavía es posible mencionar otros estadísticos de tendencia central que, aunque escasamente utilizados en el contexto de las ciencias del comportamiento y de la salud, pueden encontrarse en otros ámbitos. Nos referimos a las medias armónica y geométrica. La **media armónica** suele utilizarse con distribuciones muy asimétricas. Se obtiene dividiendo el tamaño muestral, *n*, entre la suma de los valores inversos (no tiene sentido utilizarla si algún valor de *Y* es igual a cero, pues la división por cero no está definida en el campo de los números reales):

$$\bar{Y}_{\text{arm}} = \frac{n}{\sum Y_i^{-1}}$$

La **media geométrica** suele utilizarse para promediar proporciones. Se obtiene elevando a la enésima raíz la suma de los productos de los valores absolutos de la variable (no tiene sentido utilizarla si algún valor de *Y* es negativo o cero):

$$\bar{Y}_{\text{geom}} = \left[\prod Y_i\right]^{\frac{1}{n}}$$

Promediando los logaritmos de los valores de la variable se obtiene el logaritmo de la media geométrica.

esta modificación, lógicamente, implica una pérdida de información que no puede pasarse por alto a la ligera pues, ocasionalmente, se podría estar desechando información relevante. La *mediana* lleva al límite esa modificación del tratamiento que se da a los valores más extremos: elimina del análisis todos los casos menos el(los) que ocupa(n) la posición central; de este modo, su resistencia a la presencia de anomalías en los datos es máxima. Y los estimadores *M* están a medio camino entre la modificación que aplican las medias recortada y winsorizada y la que aplica la mediana: todos ellos aprovechan las propiedades cuantitativas de los datos pero dando menos peso a los valores que más se alejan del centro; aunque no eliminan del análisis tantos casos como la mediana, asignan pesos muy pequeños a los que se alejan mucho del centro.

Los resultados de la Tabla 4.3 ilustran el comportamiento de todos estos estadísticos de tendencia central. Están calculados con los cuatro grupos de edades de la Tabla 4.2. Las distribuciones de las edades de los grupos 1 y 3 son razonablemente simétricas; las distribuciones de los grupos 2 y 4 son sensiblemente asimétricas (más adelante, en este mismo capítulo, se estudian detenidamente estos conceptos). Puede comprobarse que el grado de parecido entre los diferentes estadísticos depende del grado de asimetría de la distribución. En los grupos 1 y 3, que tienen distribuciones aproximadamente simétricas, los valores obtenidos oscilan entre 24,82 y 25,09 en el grupo 1 (una diferencia de 0,27 años) y entre 26,00 y 26,14 en el grupo 3 (una diferencia de 0,14 años). En los grupos 2 y 4, que tienen distribuciones sensiblemente asimétricas, los valores obtenidos oscilan entre 25,45 y 30,00 en el grupo 2 (una diferencia de 4,55 años) y entre 29,63 y 34,21 en el grupo 4 (una diferencia de 4,58 años). Puede comprobarse que, en estos grupos, la media aritmética es el estadístico más sensible a la presencia de asimetría en la distribución (es, de todos ellos, el que toma el valor más alto).

Tabla 4.3. Estadísticos de tendencia central aplicados a los datos de la Tabla 4.2

Estadísticos	*Grupo* 1	*Grupo* 2	*Grupo* 3	*Grupo* 4
Media aritmética	25,09	30,00	26,14	34,21
Mediana	25,00	26,00	26,00	31,00
Media recortada (5%)	24,99	28,94	26,05	32,90
Media winsorizada (5%)	24,82	28,73	26,00	33,29
Trimedia	25,00	26,75	26,00	31,25
Estimador *M* de Andrews	24,92	25,46	26,01	29,64
Estimador *M* de Hampel	24,90	26,63	26,00	31,23
Estimador *M* de Huber	25,00	26,71	26,00	30,98
Estimador *M* de Tukey	24,92	25,45	26,01	29,63

Así las cosas, ¿qué estadístico de tendencia central conviene elegir para informar de nuestros resultados? Para responder a esta pregunta hay que sopesar varios argumentos.

El primero de ellos tiene que ver con los objetivos del estudio. La fase descriptiva, con frecuencia, sólo representa el comienzo del análisis. Lo habitual es que, terminada la fase descriptiva, se pase a la inferencial para, entre otras cosas, efectuar comparaciones. Y, según veremos, las herramientas disponibles para efectuar comparaciones se basan, sobre todo, en la media aritmética. Por tanto, la media aritmética debe incluirse en el informe descriptivo porque, normalmente, hará falta para después: aunque desde el punto de vista descriptivo posee algunas limitaciones, sus propiedades inferenciales son excelentes.

Pero ya hemos señalado que la media es un estadístico demasiado sensible a la presencia de anomalías en los datos, lo cual nos lleva al segundo argumento: el centro de una variable hay que intentar identificarlo mediante estadísticos resistentes (ver Wilcox y Keselman,2003). Aunque en condiciones de simetría todos los estadísticos de tendencia central son equivalentes, en condiciones de asimetría los estadísticos resistentes (las medias modificadas, la mediana, los estimadores M) son preferibles a los no resistentes (la media). Ahora bien, como la media y la mediana es deseable que figuren en cualquier informe descriptivo porque son los estadísticos más conocidos y utilizados, el resto de estadísticos de tendencia central podrían utilizarse para recomendar a cuál de los dos, media o mediana, conviene prestar atención para identificar el centro de la distribución (en el caso de que tomen valores distintos).

Dispersión

Describir un conjunto de datos a partir de un solo número conlleva, obviamente, una importante pérdida de información: un estadístico de tendencia central informa sobre el centro de la distribución pero no dice nada sobre el resto de los valores. La consecuencia de esta limitación es que un mismo valor puede ser el centro de conjuntos de datos muy diferentes.

Los datos de la Tabla 4.4 pueden ayudar a entender esta idea. Se trata de las edades de tres grupos de sujetos. Aunque los tres grupos tienen el mismo centro (tanto la media como el resto de estadísticos de tendencia central valen 50), el grado de parecido (o el grado de *dispersión*) entre las edades del mismo grupo es muy distinto. En el primer grupo, todas las edades están muy próximas al centro: se trata de una variable con muy poca dispersión (existe mucho parecido u homogeneidad entre las puntuaciones). En el segundo grupo, la mayoría de las edades son distintas del centro pero se encuentran próximas a él: existe más dispersión que en el primer grupo pero no parece ser muy alta. En el tercer grupo, la mayoría de las edades son distintas del centro y además se encuentran muy alejadas de él: la dispersión en este grupo es mucho mayor que en los otros dos.

Tabla 4.4. Tres grupos de edades con el mismo centro y distinta dispersión

Grupos	*Edades*									
1	49	49	49	49	50	50	51	51	51	51
2	42	44	46	48	50	50	52	54	56	58
3	10	20	30	40	50	50	60	70	80	90

Lo que interesa destacar de esta discusión es que, a pesar de las evidentes diferencias existentes entre estos tres grupos de edades, conocer únicamente la edad promedio de cada grupo no permite identificar el grupo del que se está hablando (pues todos tienen la misma edad promedio): es necesario conocer, además, el grado de *dispersión*, es decir, *el grado de parecido entre los datos en el sentido de concentración o alejamiento entre ellos*.

Además, conocer el grado de dispersión de un conjunto de datos permite precisar si el centro de una distribución es o no un buen representante del resto de valores. Cuando la mayoría de los valores se encuentran próximos al centro, la dispersión es baja; cuando la mayoría de los valores se encuentran alejados del centro, la dispersión es alta. Consecuentemente, el grado de representatividad del centro de una distribución será tanto mayor cuanto menor sea

la dispersión de la distribución. Es claro, por tanto, que el grado de dispersión puede contribuir de forma importante a describir un conjunto de datos.

Ahora bien, aunque con pocos datos (como ocurre, por ejemplo, en la Tabla 4.4), una inspección visual de los mismos podría servir para formarse una idea sobre el grado de dispersión, con más datos una inspección visual resulta, además de subjetiva, completamente insuficiente. En estos casos es necesario recurrir a algún resumen capaz de cuantificar el grado de dispersión. Pero, al igual que los estadísticos de tendencia central, los de dispersión también son variados: unos utilizan todos los valores de la variable, otros sólo unos pocos; unos toman como referencia el centro de la variable, otros prescinden de él; los que toman como referencia el centro de la variable difieren en el centro que eligen; etc.

En los apartados que siguen se han agrupado estos estadísticos en tres bloques atendiendo al criterio de dispersión en el que se basan.

Amplitudes

El estadístico de dispersión más simple de todos consiste en calcular la diferencia entre el valor más grande ($Y_{\text{máx}}$) y el más pequeño ($Y_{\text{mín}}$). A esta diferencia se le llama **amplitud total** (A_T), aunque también puede encontrarse como *rango*, *recorrido* o, simplemente, *amplitud*:

$$A_T = Y_{\text{máx}} - Y_{\text{mín}} \qquad [4.8]$$

Calculada en los tres grupos de edades de la Tabla 4.4, la amplitud toma los siguientes valores: $A_1 = 51 - 49 = 2$; $A_2 = 58 - 42 = 16$; $A_3 = 90 - 10 = 80$. Estos valores parecen reflejar la dispersión baja (2 años), media (16 años) y alta (80 años) que anteriormente hemos intuido tras una inspección visual de los datos. Pero no hay que olvidar que en este estadístico únicamente intervienen los dos valores extremos, lo cual significa que no se está prestando atención a la disposición de los valores intermedios. Y esto tiene dos consecuencias indeseables. En primer lugar, puede ocurrir que conjuntos de datos muy diferentes tengan la misma amplitud total; basta con que los dos valores extremos sean los mismos. En segundo lugar, la presencia de un solo caso muy distante del resto es capaz de alterar sensiblemente el valor de la amplitud total (por muy parecidos entre sí que sean el resto de los valores). Se trata, por tanto, de un estadístico muy poco resistente. No obstante, al estar expresado en la misma métrica que la variable, su interpretación es muy sencilla. Y con distribuciones simétricas permite formarse una idea muy rápida (y podríamos decir, también, acertada) del grado de dispersión.

Otro estadístico basado en la misma lógica que la amplitud total, pero más resistente que ésta, es la **amplitud intercuartil** (A_{IQ}). Se basa en la distancia existente entre los cuartiles primero y tercero: $A_{IQ} = Q_3 - Q_1$ (ya hemos hablado de los cuartiles en este mismo capítulo, en el apartado *Cuantiles*). La amplitud intercuartil mide el grado de dispersión del 50% de los casos centrales (normalmente, los más representativos del conjunto de datos) y, de esta manera, resuelve los problemas derivados de prestar atención únicamente a los dos casos más extremos; pero no puede pasarse por alto el hecho de que, al prestar atención únicamente al 50% de los casos centrales, se puede estar desechando información relevante. Más adelante, en este mismo capítulo, veremos que uno de los gráficos más interesantes para describir la forma y otras propiedades de una distribución (el *diagrama de caja*) se basa justamente en este estadístico. En ocasiones se utiliza otro estadístico llamado *amplitud semi-intercuartil* (la amplitud intercuartil dividida entre 2), pero su significado es menos claro que el de la amplitud intercuartil (aunque, lógicamente, es algo más resistente).

Desviaciones promedio

Además de los estadísticos basados en el concepto de distancia entre dos valores, existen otros que, no sólo permiten expresar el grado de dispersión en la misma métrica de la variable (como hacen las amplitudes), sino que lo hacen utilizando *todos* los valores.

Es claro que la desviación que una puntuación experimenta respecto del centro (media) de su distribución está expresando el grado de dispersión de esa puntuación. Para resaltar esta idea, la Tabla 4.5 ofrece las desviaciones (puntuaciones diferenciales: $y = Y - \bar{Y}$) correspondientes a los tres grupos de edades de la Tabla 4.4. Sabemos por la *amplitud total* (y por la simple inspección visual de los datos) que las edades del segundo grupo están más dispersas que las del primero; también sabemos que las edades del tercer grupo están más dispersas que las de los otros dos grupos. Esta diferencia en las dispersiones queda claramente reflejada en las puntuaciones diferenciales recogidas en la Tabla 4.5: las edades del grupo 1 toman valores entre −1 y 1; las del grupo 2 entre −8 y 8; y las del grupo 3 entre −40 y 40.

Tabla 4.5. Puntuaciones diferenciales correspondientes a las edades de la Tabla 4.4

Grupos	*Puntuaciones diferenciales* ($y_i = Y_i - \bar{Y}$)										$\Sigma\|y_i\|$	Σy_i^2
1	−1	−1	−1	−1	0	0	1	1	1	1	8	8
2	−8	−6	−4	−2	0	0	2	4	6	8	40	240
3	−40	−30	−20	−10	0	0	10	20	30	40	200	6.000

Ahora bien, si la desviación que una puntuación individual experimenta respecto de la media de su distribución está expresando el grado de dispersión de esa puntuación, la suma de todas las desviaciones de un conjunto de datos debería estar expresando la dispersión total del conjunto. Sin embargo, sabemos que esto no es así porque la suma de las desviaciones de la media vale cero (recordemos que las desviaciones positivas se anulan con las negativas). Para evitar este problema basta con aplicar algún tipo de transformación que haga tomar a todas las desviaciones un valor positivo (de hecho, puesto que una distancia entre dos elementos no puede ser negativa, lo recomendable, desde el punto de vista de la dispersión, es expresar las distancias a la media en valor positivo). Las transformaciones más habituales consisten en tomar el valor absoluto de las desviaciones o en elevarlas al cuadrado.

En la **media de las desviaciones** ($\bar{Y}_{|\text{desv}|}$), también llamada *desviación media* y *desviación media absoluta*, las desviaciones se toman en valor absoluto y su suma se divide por el número de puntuaciones para obtener la desviación promedio:

$$\bar{Y}_{|\text{desv}|} = \frac{\sum |Y_i - \bar{Y}|}{n} = \frac{\sum |y_i|}{n} \qquad [4.9]$$

La *media de las desviaciones* recoge la idea de dispersión respecto del valor central (la media) y representa el promedio del conjunto de distancias a la media. Calculada sobre los tres grupos de edades de la Tabla 4.4 (la Tabla 4.5 ofrece la suma de los valores absolutos de las puntuaciones diferenciales) se obtienen los siguientes resultados:

$\bar{Y}_{|\text{desv}|\,(1)} = 8/10 = 0{,}8$

$\bar{Y}_{|\text{desv}|\,(2)} = 40/10 = 4$

$\bar{Y}_{|\text{desv}|\,(3)} = 200/10 = 20$

Puesto que la *media de las desviaciones* está expresada en la misma métrica que la variable original, los valores obtenidos reflejan la dispersión de cada grupo en unidades de *edad* (es decir, en años). Por tanto, las edades del primer grupo se desvían de su media, en promedio, algo menos de un año; las del segundo grupo, 4 años; y las del tercer grupo, 20 años.

Estudiar la dispersión a partir de las desviaciones del centro no implica tener que trabajar necesariamente con la media aritmética. Debe tenerse en cuenta que la media de las desviaciones sigue siendo una media y que, como tal, adolece de excesiva sensibilidad a la presencia de valores extremos por una de las dos colas de la distribución (asimetría). Este problema puede resolverse utilizando estadísticos resistentes como la *mediana*. Esto es precisamente lo que hace un estadístico denominado **mediana de las desviaciones** ($Mdn_{|\text{desv}|}$) o *mediana de las desviaciones absolutas*:

$$Mdn_{|\text{desv}|} = Mdn\,|Y_i - Mdn_Y| \qquad [4.10]$$

La *mediana de las desviaciones* es, por tanto, la mediana de las distancias a la mediana. Si la distribución es simétrica, la *mediana* de las desviaciones tomará un valor parecido al de la *media* de las desviaciones (puesto que la media y la mediana de una distribución simétrica son iguales, también lo serán las distancias a la media y a la mediana). Pero conforme la distribución se vaya haciendo más asimétrica, el valor de la *media* de las desviaciones se verá más alterado que el de la *mediana* de las desviaciones. Calculando la *mediana de las desviaciones* en los tres grupos de edades de la Tabla 4.4 se obtiene:

$$Mdn_{|\text{desv}|\,(1)} = Mdn\,[0,\ 0,\ 1,\ 1,\ 1,\ 1,\ 1,\ 1,\ 1,\ 1] = 1$$
$$Mdn_{|\text{desv}|\,(2)} = Mdn\,[0,\ 0,\ 2,\ 2,\ 4,\ 4,\ 6,\ 6,\ 8,\ 8] = 4$$
$$Mdn_{|\text{desv}|\,(2)} = Mdn\,[0,\ 0,\ 10,\ 10,\ 20,\ 20,\ 30,\ 30,\ 40,\ 40] = 20$$

Puesto que las tres distribuciones de edad son simétricas, los valores obtenidos con la *mediana* de las desviaciones son casi idénticos a los obtenidos con la *media* de las desviaciones. Sin embargo, introduciendo un dato anómalo en el segundo grupo (cambiando, por ejemplo, 58 por 90), la *media* de las desviaciones pasa de 4 a 8,08, mientras que la *mediana* de las desviaciones no se altera. Además, como la *mediana* de las desviaciones se encuentra, al igual que la *media* de las desviaciones, en la misma métrica que la variable original, los valores obtenidos reflejan la dispersión en la misma métrica de la variable (años).

Varianza y desviación típica

La *media de las desviaciones* y la *mediana de las desviaciones* no son los únicos estadísticos que intentan cuantificar la dispersión a partir de las desviaciones del centro de la distribución. Puesto que esas desviaciones representan la esencia del concepto de dispersión, es lógico que se les haya prestado especial atención. En este apartado se estudian dos de los estadísticos de dispersión más utilizados. Ambos se basan en las desviaciones de la media, pero, para evitar que su suma valga cero, en lugar de tomar esas desviaciones en valor absoluto (estrategia utilizada por la *media de las desviaciones*), se elevan al cuadrado.

La **varianza** (Fisher, 1918) es el promedio de las desviaciones cuadráticas de la media, es decir, el promedio de las desviaciones de la media elevadas al cuadrado. Se trata, por tanto, de una *media de cuadrados* (o, como suele denominarse en algunos contextos, una *media cuadrática*). Se define y representa mediante

$$S^2_{Y(n)} = \frac{\sum (Y_i - \bar{Y})^2}{n} \qquad [4.11]$$

Puesto que el numerador de la varianza recoge las desviaciones de la media (es decir, las puntuaciones *diferenciales*; ver ecuación [4.3]), la varianza también puede expresarse como

$$S^2_{Y(n)} = \frac{\sum y_i^2}{n} \qquad [4.12]$$

El subíndice Y identifica a la variable; el subíndice n indica el valor del denominador. Prestar atención explícita al valor del denominador tiene su explicación. Ocurre que, cuando se utiliza la varianza muestral para hacer inferencias sobre la varianza poblacional (trataremos esta cuestión en el Capítulo 7), utilizar n en el denominador da como resultado un valor muestral que tiende a ser menor que el valor poblacional. Este sesgo desaparece simplemente restando 1 al tamaño muestral en el denominador. A esta modificación de la varianza se le suele llamar **varianza insesgada** (también puede encontrarse con el nombre de *cuasivarianza*):

$$S^2_{Y(n-1)} = S^2_Y = \frac{\sum (Y_i - \bar{Y})^2}{n-1} = \frac{\sum y_i^2}{n-1} \qquad [4.13]$$

Cuando se utiliza un programa informático como el SPSS para calcular la varianza, el resultado que se obtiene es, siempre, la varianza insesgada. Por tanto, para evitar confusiones, a partir de ahora únicamente haremos referencia a la varianza insesgada, lo cual permitirá identificarla simplemente como S^2_Y.

Aplicando la ecuación [4.13] a los tres grupos de edades de la Tabla 4.4 (la Tabla 4.5 recoge las puntuaciones diferenciales y la suma de sus cuadrados) se obtiene:

$$S^2_{Y|\,\text{grupo 1}} = [(-1)^2 + (-1)^2 + \cdots + (1)^2]/(10-1) = 8/9 = 0{,}89$$
$$S^2_{Y|\,\text{grupo 2}} = [(-8)^2 + (-6)^2 + \cdots + (8)^2]/(10-1) = 240/9 = 26{,}67$$
$$S^2_{Y|\,\text{grupo 3}} = [(-40)^2 + (-30)^2 + \cdots + (40)^2]/(10-1) = 6.000/9 = 666{,}67$$

Probablemente, la primera reacción que se tiene al observar el valor de la varianza es preguntarse por su significado. El valor obtenido en el primer grupo se parece al obtenido con la desviación media (0,89 frente a 0,8); pero esto es sólo porque, en ese grupo, las distancias a la media no superan el valor 1. En el segundo grupo se ha obtenido un valor claramente mayor que el obtenido con la desviación media (26,67 frente a 4). Y en el tercer grupo se ha obtenido un valor muchísimo mayor (666,67 frente a 20). Ahora bien, si se tiene en cuenta que las edades del segundo grupo oscilan entre 42 y 58 (16 años), no parece que una dispersión de 26,67 esté reflejando algo real. Y menos real todavía se antoja el resultado obtenido en el tercer grupo: 666,67 es un número que no tiene nada que ver con la edad.

Lo cierto es que la varianza no permite formarse una idea acertada del grado de dispersión de una variable. La razón de esto es que las distancias a la media están elevadas al cuadrado: es como intentar medir la distancia física entre dos puntos utilizando una medida de superficie en lugar de una medida de longitud. La varianza puede servir para comparar entre sí distintos grupos (lógicamente, en la misma variable) y saber en cuál de ellos hay mayor dispersión, pero no sirve para formarse una idea sobre el grado de dispersión. Por esta razón, lo que suele hacerse es utilizar la raíz cuadrada de la varianza. Es decir,

$$S_Y = \sqrt{\sum (Y_i - \bar{Y})^2/(n-1)} = \sqrt{\sum y_i^2/(n-1)} \qquad [4.14]$$

(siempre se toma la raíz positiva, pues la dispersión puede ser nula o positiva, pero no negativa). A este valor se le llama **desviación típica** o *desviación estándar* (Pearson, 1894). Como en el denominador se está utilizando $n-1$, en realidad se trata de la desviación típica *insesgada*. Y como se trata de la raíz cuadrada de valores previamente elevados al cuadrado, el resultado está más en consonancia con la métrica de la variable cuya dispersión se intenta cuantificar. Aplicando [4.14] a los tres grupos de edades de la Tabla 4.4 se obtienen las raíces cuadradas de los valores ya obtenidos con [4.13]:

$$S_{Y|\,\text{grupo 1}} = \sqrt{0{,}89} = 0{,}94$$
$$S_{Y|\,\text{grupo 2}} = \sqrt{26{,}67} = 5{,}16$$
$$S_{Y|\,\text{grupo 3}} = \sqrt{666{,}67} = 25{,}82$$

Al igual que ocurre con la media de las desviaciones, la desviación típica es un promedio basado en las desviaciones de la media. Cabría esperar, por tanto, que el valor de la desviación típica fuera parecido al de la media de las desviaciones; pero lo cierto es que esto sólo ocurre cuando se trabaja con distribuciones simétricas. La consecuencia de elevar al cuadrado las desviaciones de la media es que las desviaciones más grandes reciben mayor peso que las pequeñas (aunque después se tome la raíz cuadrada del promedio de esas desviaciones cuadráticas, la desigual ponderación de las desviaciones más grandes ya está hecha y queda reflejada en el resultado). Por tanto, la desviación típica tomará un valor tanto más alejado del de la media de las desviaciones cuanto más asimétrica sea la distribución. No obstante, pensar en la desviación típica como en un *promedio de distancias a la media* permite formarse una idea sobre su significado sin distorsionarlo seriamente.

Comparación entre estadísticos de dispersión

La dispersión es un *concepto esencialmente positivo*: o todos los valores de la variable son iguales y, consecuentemente, no existe dispersión (en cuyo caso no estaríamos hablando de una *variable* sino de una *constante*), o unos valores son distintos de otros y, consecuentemente, existe dispersión (en cuyo caso hay que cuantificar si es baja, media o alta; pero nunca negativa). Esta idea es lo bastante importante como para no pasarla por alto, pero no ayuda a elegir un estadístico de dispersión: todos ellos valen cero cuando no hay dispersión y toman un valor positivo cuando sí la hay.

Otro aspecto de la dispersión que conviene no pasar por alto es la forma en que cambia cuando se efectúan transformaciones en los datos. En relación con esto es importante señalar que la *dispersión no se altera cuando a los valores de una variable se les suma o resta una constante*. Imaginemos que las edades de las que venimos hablando (Tabla 4.4) vuelven a registrarse pasados 3 años. Todos los sujetos tendrán 3 años más (es decir, la edad media habrá aumentado 3 años), pero la dispersión entre las edades seguirá siendo la misma porque el grado de concentración o alejamiento entre ellas no se habrá alterado. De nuevo se trata de un aspecto importante, pero que no permite distinguir entre los estadísticos propuestos: todos ellos permanecen inalterados cuando a una variable se le suma una constante.

Pero, ¿qué ocurre cuando, en lugar de sumar una constante, se multiplica o divide[7]? Lo que ocurre es que *la dispersión queda multiplicada o dividida por esa constante*. Por ejemplo, ¿qué ocurre si todas las edades se multiplican por 3 años? Pues que la dispersión se triplica. Si se multiplican por 3 las edades del primer grupo, la edad más baja (49) pasa a 147 y la más alta (51) pasa a 153. Con ello, la amplitud inicial de 2 años aumenta hasta 6 años. Pues bien, todos los estadísticos estudiados, excepto la varianza, reflejan correctamente este cambio en la dispersión; es decir, todos ellos, excepto la varianza, triplican su valor cuando los valores originales de la variable se multiplican por 3 (lo que ocurre con la varianza es que queda multiplicada por el cuadrado de la constante; por tanto, estaría indicando que la dispersión ha aumentado 9 veces en lugar de 3).

Otro aspecto más al que conviene prestar atención tiene que ver con *la métrica en la que viene expresado cada estadístico*. Los estadísticos cuyos valores se encuentran en la misma métrica que la variable original tienen la ventaja de que son más fáciles de interpretar; por tanto, son más útiles para formarse una idea sobre el grado de dispersión de una variable. Y ocurre que, de nuevo, a excepción de la varianza, todos los estadísticos estudiados cuantifican la dispersión en la misma métrica que la variable original (salvando las matizaciones ya hechas sobre la desviación típica).

Parece, por tanto, que la varianza no se encuentra entre los estadísticos que conviene elegir para cuantificar la dispersión: no sólo no se encuentra en la misma métrica de la variable original, sino que no refleja correctamente el cambio que experimenta la dispersión cuando se aplican algunas transformaciones simples a los datos[8].

Ahora bien, descartada la varianza, ¿cómo elegir entre el resto de estadísticos? Al igual que ocurre con las medidas de tendencia central, la clave se encuentra de nuevo en la *forma de la distribución*. Ya sabemos que las amplitudes (total, intercuartil) se basan en la distancia entre sólo dos valores; también sabemos que las desviaciones promedio (media de las desviaciones, mediana de las desviaciones y desviación típica) se basan en las distancias al centro de la distribución. Pues bien, el hecho mismo de disponer de estadísticos basados en criterios tan dispares ya parece estar recomendando que siempre se informe con uno de cada tipo: una *amplitud* y una *desviación*.

Pero, ¿cómo elegir entre las *amplitudes*? Ya hemos señalado que la amplitud total adolece de serios inconvenientes: puede tomar el mismo valor con conjuntos de datos muy distintos y es extremadamente sensible a la presencia de casos anómalos (un sólo caso anómalo basta para alterar por completo su valor). Por tanto, la amplitud total sólo debería utilizarse con distribuciones simétricas o aproximadamente simétricas. Si no se da esta circunstancia, es preferible utilizar la amplitud intercuartil. Pero sin olvidar que la amplitud intercuartil desecha una parte importante de los datos (el 25% de cada lado de la distribución) y que también esto tiene sus inconvenientes. Por supuesto, saber qué interesa conocer exactamente de una variable puede ayudar a elegir entre ambas amplitudes; pero, independientemente de dónde se sitúe el interés descriptivo, las amplitudes poseen importantes debilidades que no deben pasarse por alto.

---

[7] Multiplicar o dividir una variable por una constante es una tarea relativamente frecuente. Por ejemplo, si una prueba diagnóstica arroja puntuaciones de 0 a 20 y se prefiere transformarlas en una métrica de 0 a 100 (simplemente porque se considera una métrica más fácil de entender), bastará con multiplicar las puntuaciones originales por 5.

[8] A pesar de sus evidentes limitaciones desde el punto de vista descriptivo, lo cierto es que la varianza (o, mejor, la varianza insesgada) posee, según tendremos ocasión de comprobar, excelentes propiedades inferenciales que la convierten en un estadístico esencial en el contexto de la estimación de parámetros y del contraste de hipótesis.

Y ¿cómo elegir entre las *desviaciones*? Ya hemos visto que los estadísticos que se basan en las desviaciones del centro no son igualmente sensibles al grado de asimetría de la distribución. Aunque en una distribución simétrica todos ellos toman valores parecidos, en presencia de asimetría la desviación típica es menos resistente que la media de las desviaciones y ésta menos resistente que la mediana de las desviaciones. Elegir el estadístico más resistente, por tanto, no parece una tarea complicada. Pero ocurre que la desviación típica, a pesar de ser la desviación menos resistente, posee una serie de propiedades que hacen de ella un estadístico especialmente interesante. Entre otras cosas, la desviación típica constituye la base de unas transformaciones llamadas *puntuaciones típicas* que poseen una extraordinaria utilidad para conocer, por ejemplo, la posición relativa que ocupa cada caso dentro de su grupo (se estudian en el próximo capítulo). Además, cuando se tiene intención de pasar de la fase descriptiva a la inferencial para efectuar comparaciones y estudiar relaciones, la desviación típica adquiere (tendremos ocasión de comprobarlo) un protagonismo que no puede ser asumido por el resto de los estadísticos de dispersión. Por tanto, por los mismos argumentos que, según hemos visto ya, es tarea obligada informar del valor de la media aritmética a pesar de su baja resistencia, también lo es informar de la desviación típica (esta razón, unida al hecho de que ambos estadísticos están expresados en la misma métrica, es la que justifica la práctica habitual de ofrecer juntas la media y la desviación típica en los informes de resultados).

La información recogida en la Tabla 4.6 puede ayudar a aclarar el comportamiento de todos estos estadísticos. Están calculados con las edades de la Tabla 4.2. Recordemos que las edades de los grupos 1 y 3 se distribuyen de forma aproximadamente simétrica, mientras que las de los grupos 2 y 4 se distribuyen de forma sensiblemente asimétrica. Debido a que los estadísticos propuestos se basan en diferentes criterios de dispersión, los valores que toman son muy distintos. Por lo que se refiere a las *amplitudes*, en condiciones de simetría y poca dispersión (grupos 1 y 3), la amplitud total es aproximadamente el doble de la intercuartil. Sin embargo, en condiciones de asimetría y mayor dispersión (grupos 2 y 3), la amplitud total alcanza aproximadamente el triple de la intercuartil. Con las *desviaciones* ocurre algo parecido. En condiciones de simetría (grupos 1 y 3), las tres desviaciones toman valores parecidos: oscilan entre 3 y 4,11 en el grupo 1 (una diferencia de 1,11 años) y entre 4 y 4,90 en el grupo 3 (una deferencia de 0,90 años), siendo la desviación típica la que toma los valores más altos. En condiciones de asimetría (grupos 2 y 4), los valores de las distintas desviaciones cambia de forma notable: oscilan entre 6 y 13,02 en el grupo 2 (una diferencia de 7,02 años) y entre 9,50 y 16,10 en el grupo 4 (una diferencia de 6,60 años). El estadístico más resistente es la mediana de las desviaciones; el menos resistente, la desviación típica.

Tabla 4.6. Estadísticos de dispersión aplicados a los datos de la Tabla 4.2

Estadísticos	*Grupo* 1	*Grupo* 2	*Grupo* 3	*Grupo* 4
Amplitud total	14	43	16	56
Amplitud intercuartil	6	15	8,5	20
Media de las desviaciones	3,19	9,45	4,14	11,82
Mediana de las desviaciones	3	6	4	9,5
Desviación típica	4,11	13,02	4,9	16,1
Varianza	16,89	169,4	23,98	259,26
Coeficiente de variación (media)	16,38	43,38	18,73	47,06
Coeficiente de variación (mediana)	16,44	52,6	18,84	60,48

Coeficientes de variación

Cuantificar la dispersión no sólo es útil para conocer el grado de dispersión de una variable (con lo que ello implica de caracterización de la variable, valoración del grado de representatividad del centro de la distribución, etc.), sino también para comparar la dispersión de diferentes grupos o variables. Los estadísticos de dispersión estudiados permiten hacer ambas cosas, pero con sus limitaciones.

Las desviaciones típicas de la Tabla 4.6 indican que las edades de los grupos 1, 2, 3 y 4 se distancian de sus respectivos centros, en promedio, en torno a 4, 13, 5 y 16 años, respectivamente. ¿Permite esto afirmar que el grupo 1 es el menos disperso y el grupo 4 el más disperso? En principio, sí, pues todas las medidas de dispersión coinciden en señalar ese hecho y los cuatro grupos de puntuaciones se refieren a la misma variable (la edad). Pero cuando se compara la dispersión de distintos grupos debe tenerse en cuenta la magnitud de los valores que se comparan, pues con valores pequeños cabe esperar encontrar menos dispersión que con valores grandes. Esto se comprenderá fácilmente si se considera la edad medida en meses y en años; los valores expresados en meses serán mucho más grandes que los expresados en años (12 veces más grandes) y las desviaciones del promedio en meses serán mucho mayores que las del promedio en años. Además, cuando se compara la dispersión de variables distintas es importante prestar atención a su métrica (es decir, a las unidades de medida utilizadas). Al comparar la dispersión de las alturas y los pesos de un determinado grupo de sujetos, no podrá pasarse por alto el hecho de que se están comparando *cm* con *kg*.

Para facilitar la *comparación entre grupos y variables* se han diseñado estadísticos de dispersión *relativa*. El más utilizado de éstos, el **coeficiente de variación centrado en la media** (CV_{Media}), expresa la desviación típica como un porcentaje del valor absoluto de la media:

$$CV_{\text{Media}} = \frac{S_Y}{|\bar{Y}|}\,100 \qquad [4.15]$$

El **coeficiente de variación centrado en la mediana** (CV_{Mediana}) se obtiene de la siguiente manera: se calculan las desviaciones de la mediana, se elevan al cuadrado, se promedian utilizando $n-1$ (se tiene así una especie de varianza basada en la mediana), se obtiene la raíz cuadrada de ese promedio y el resultado se divide entre el valor absoluto de la mediana. Todo ello se multiplica por 100 para expresar el resultado final como un porcentaje:

$$CV_{\text{Mediana}} = \frac{S_{\text{Mdn}}}{|Mdn|}\,100 \qquad [4.16]$$

Aunque con distribuciones simétricas o aproximadamente simétricas ambos coeficientes ofrecen resultados parecidos (ver Tabla 4.6, grupos 1 y 3), en distribuciones asimétricas pueden ofrecer resultados muy distintos (ver Tabla 4.6, grupos 2 y 4). Si los valores más alejados del centro se encuentran en la parte alta de la distribución, el coeficiente centrado en la mediana será mayor (incluso mucho mayor) que el centrado en la media; si los valores más alejados del centro se encuentran en la parte baja de la distribución, el coeficiente centrado en la mediana será menor (incluso mucho menor) que el centrado en la media.

Una dispersión razonable va asociada a coeficientes de variación menores que 50. Coeficientes de variación mayores de 50 indican mucha dispersión. Coeficientes mayores que 100 están delatando, generalmente, fuertes anomalías en los datos.

Forma de la distribución

Al reflexionar acerca de las fortalezas y debilidades de los diferentes estadísticos de tendencia central y de dispersión nos hemos visto obligados a hacer constantes referencias a la forma de la distribución (particularmente al grado de asimetría). Pero la forma de la distribución, además de ayudar a elegir entre estadísticos, es útil por sí misma: permite obtener una visión rápida de las características de la variable, detectar valores anómalos (valores que no se parecen al resto), advertir inconsistencias en los datos (valores que se repiten demasiado o valores que no aparecen), etc. Por tanto, conocer la forma de una distribución tiene un interés comparable al de identificar su centro o cuantificar su dispersión.

Estudiar la forma de una distribución implica, básicamente, valorar dos características: *asimetría* y *curtosis*. La **asimetría** se refiere a la forma en que los datos se distribuyen por encima y por debajo del centro. En una distribución simétrica, las distribuciones a cada lado del centro tienen la misma forma (una es espejo de la otra). La simetría se rompe cuando existen casos que se alejan del centro más por uno de los extremos que por el otro. Cuando los casos más alejados del centro se encuentran en la zona alta de la distribución, decimos que existe *asimetría positiva*; cuando se encuentran en la zona baja, *asimetría negativa*.

La **curtosis** de una distribución se refiere a su grado de *apuntamiento*. Esta característica se valora por comparación con una distribución teórica llamada *curva normal* que se estudia en el siguiente capítulo. Esta distribución teórica se toma como referente de curtosis media. La curtosis expresa el grado en que una distribución acumula casos en sus colas en comparación con los casos que acumulan las colas de una curva normal con la misma media y con la misma desviación típica. El concepto de curtosis es algo confuso (de esta confusión no se libran muchos manuales de estadística y de análisis de datos). Una distribución *leptocúrtica* es más puntiaguda que la curva normal y, además, acumula más casos que ésta en sus colas. Una distribución *platicúrtica* es más aplastada que la curva normal y, además, acumula menos casos que ésta en sus colas. A la curva normal (referente de curtosis media) se le llama *mesocúrtica*.

Para valorar estas dos características contamos con dos tipos de herramientas: *gráficos* y *estadísticos*. En los dos aparados siguientes se estudian ambas cosas.

Gráficos para variables cuantitativas

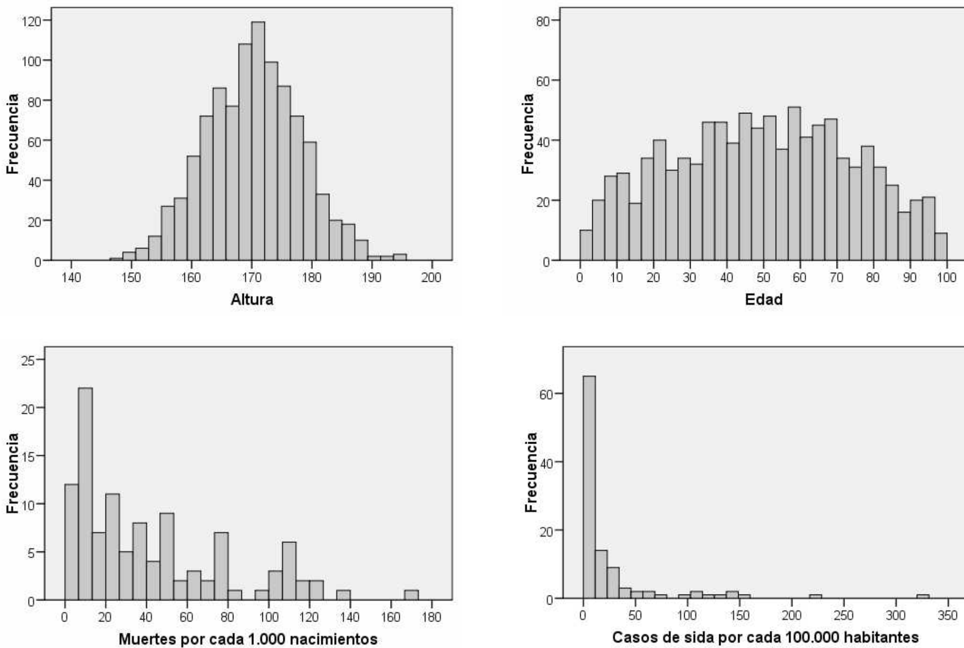
Entre los gráficos disponibles para describir la forma de la distribución de una variable cuantitativa los más utilizados son: el histograma, el polígono de frecuencias, el diagrama de tallo y hojas, y el diagrama de caja.

El **histograma** es parecido al gráfico de barras ya estudiado en el capítulo anterior, pero con las barras juntas, dando así una impresión de continuidad que no da el diagrama de barras. Se construye sobre el plano definido por dos ejes cartesianos: en el eje horizontal se colocan los valores de la variable ordenados de menor a mayor (comenzando por la izquierda), en el vertical se colocan las frecuencias (número de veces que se repite cada valor) y sobre cada valor se levanta una barra o rectángulo de altura proporcional a su frecuencia (la anchura de las barras no es relevante). Puesto que entre los valores de una variable cuantitativa se dan muy pocas repeticiones, lo habitual es formar intervalos agrupando unos pocos valores consecutivos y utilizar esos *intervalos* en el eje horizontal (en lugar de cada valor individual).

La Figura 4.1 muestra varios ejemplos de histogramas. Los dos de la mitad superior representan datos de una muestra de 997 sujetos. Ambos muestran distribuciones aproximadamente simétricas, pero con una pauta algo distinta. En el primero de ellos, las frecuencias (la altura de las barras) van disminuyendo rápidamente conforme los valores se alejan del centro de la distribución (tendremos ocasión de constatar repetidamente que este tipo de distribuciones se dan con mucha frecuencia en el mundo real); en el segundo, la disminución de las frecuencias es más lenta, dando la impresión, incluso, de que los extremos están recortados (enseguida veremos que esta pauta es típica de las distribuciones platicúrticas).

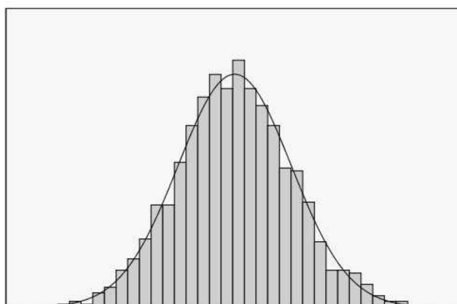
Los dos histogramas de la mitad inferior representan datos correspondientes a 109 países. Ambos muestran distribuciones con asimetría positiva: los casos tienden a concentrarse en la zona baja de la distribución o, lo que es lo mismo, los casos más alejados del centro se encuentran en la zona alta de la distribución. Pero la asimetría del segundo histograma es mucho más acusada que la del primero; en el segundo se observan claramente algunos casos muy distanciados del resto.

Figura 4.1. Histogramas de las variables *altura*, *edad*, *tasa de muertes al nacer* y *tasa de sida*



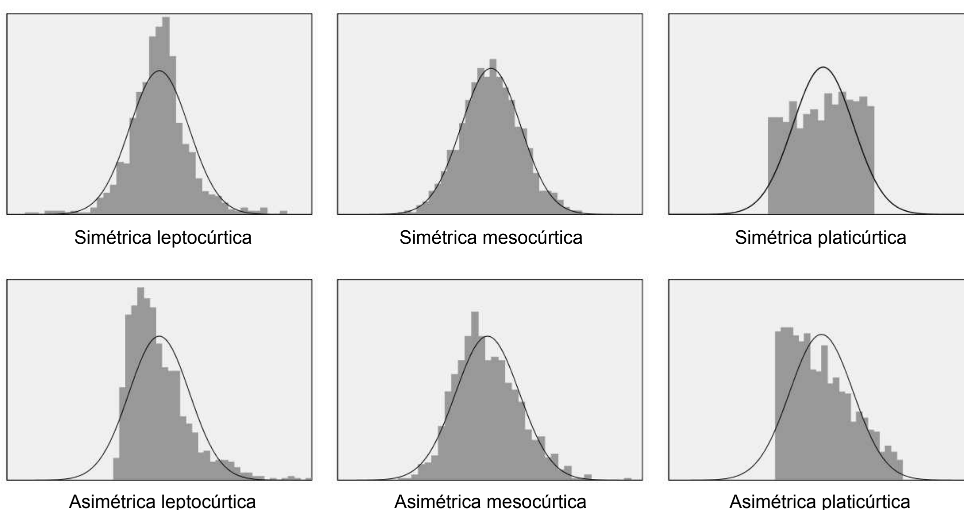
La asimetría de una distribución es más fácil de observar que la curtosis. Para hacerse una idea acerca de si una distribución es leptocúrtica o platicúrtica conviene compararla con una distribución mesocúrtica: la curva normal (aunque más adelante, en el próximo capítulo, se estudia detenidamente esta curva –ver el apartado *Curva normal*–, para los propósitos de este apartado basta con saber que se trata de una curva simétrica y mesocúrtica). La Figura 4.2 muestra el histograma de una distribución aproximadamente normal junto con una curva normal superpuesta. Lógicamente, la valoración de la curtosis es más fácil cuando se toma como referencia esta curva teórica.

Figura 4.2. Histograma con curva normal



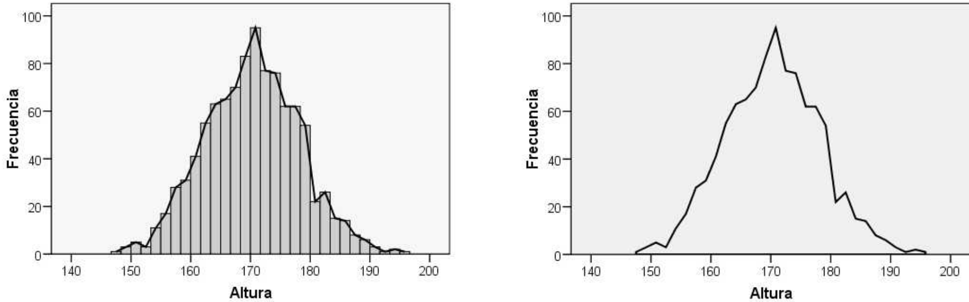
Los histogramas de la Figura 4.3 pueden ayudar a precisar el concepto de curtosis. Los tres de la mitad superior representan distribuciones simétricas con distinto grado de curtosis: el de la izquierda muestra una distribución leptocúrtica (en ambas colas de la distribución hay más casos que en las de la curva normal), el del centro muestra una distribución mesocúrtica (se ajusta bastante bien a la curva normal) y el de la derecha muestra una distribución platicúrtica (en ambas colas de la distribución hay menos casos que en las colas de la curva normal). Los tres histogramas de la mitad inferior representan distribuciones asimétricas con distinto grado de curtosis: el de la izquierda muestra una distribución leptocúrtica (por la cola derecha hay más casos que en la curva normal), el del centro muestra una distribución mesocúrtica (sus colas acumulan más o menos los mismos casos que las colas de la curva normal), y el de la derecha muestra una distribución platicúrtica (en ambas colas de la distribución hay menos casos que en las de la curva normal). Dicho esto, quizá no esté de más advertir que, de estas dos características, la simetría suele interesar más que la curtosis. Entre otras cosas, el grado de curtosis no afecta al centro de la distribución de forma tan evidente como el grado de asimetría.

Figura 4.3. Histogramas correspondientes a distribuciones con distinto grado de asimetría y curtosis



Uniendo con una línea los puntos medios de los bordes superiores de cada barra del histograma (Figura 4.4, izquierda) se obtiene el **polígono de frecuencias** (Figura 4.4, derecha). La información de este gráfico es prácticamente idéntica a la de un histograma. Según veremos, en estadística inferencial es bastante habitual representar la forma de las distribuciones mediante polígonos de frecuencias *suavizados* (como si los intervalos creados para dibujar el correspondiente histograma fueran infinitamente pequeños; ver Figura 4.2).

Figura 4.4. Histograma y polígono de frecuencias de la variable *altura*



Si las barras del histograma se sustituyen por los valores de la variable se obtiene un gráfico llamado **diagrama de tallo y hojas** (*stem and leaf plot*; Tukey, 1977) que informa simultáneamente de los valores que toma la variable y de la forma de la distribución. Ofrece la misma información que un histograma, pero más detallada: permite comprobar si existen valores que se repiten mucho o valores que no aparecen. Suele representarse en horizontal, no en vertical como el histograma. La Figura 4.5 muestra el diagrama de tallo y hojas de la variable *altura* (se corresponde con el primer histograma de la Figura 4.1 y con el polígono de frecuencias de la Figura 4.4). La Figura 4.6 muestra el diagrama de la variable *muertes por cada mil nacimientos* (se corresponde con el tercer histograma de la Figura 4.1).

Al igual que en un histograma, la longitud de las líneas refleja el número de casos que pertenecen a cada intervalo de valores (la frecuencia exacta con la que se repite cada valor viene indicada en la primera columna del gráfico). Y cada caso o grupo de casos está representado por un número que no es otra cosa que el valor de ese caso en la variable.

Cada valor se descompone en dos partes: el primer o primeros dígitos forma(n) el *tallo* (*stem*; valor inmediatamente anterior al punto) y los dígitos que siguen al tallo forman las hojas (*leaf*; valores que siguen al punto). Por ejemplo, el valor 23 se descompone en un tallo de 2 y en una hoja de 3; el valor 178 se descompone en un tallo de 17 y en una hoja de 8; etc.

Cada tallo puede ocupar una sola fila o varias. Si ocupa una sola fila, sus hojas contienen dígitos del 0 al 9; si ocupa dos filas, las hojas de la primera fila contienen dígitos del 0 al 4 y las de la segunda fila dígitos del 5 al 9; etc. Los tallos de la Figura 4.5 (excepto el primero y el último) ocupan cinco filas: la primera fila (primera hoja) contiene los dígitos 0 y 1; la segunda, los dígitos 2 y 3; la tercera, los dígitos 4 y 5; la cuarta, los dígitos 6 y 7; y la quinta, los dígitos 8 y 9. Los tallos de la Figura 4.6 ocupan una sola fila.

Cuando la anchura del tallo vale 10 (como ocurre en los ejemplos de las Figuras 4.5 y 4.6), los dígitos de las hojas son unidades; cuando la anchura del tallo vale 100, los dígitos de las hojas son decenas; cuando la anchura del tallo vale 1.000, los dígitos de las hojas son centenas; etc. La anchura del tallo (*Stem width*) se indica en la parte inferior del diagrama y es un dato imprescindible para interpretarlo correctamente. En el ejemplo de la Figura 4.5 el

tallo tiene una anchura de 10, lo que significa que los valores del tallo hay que multiplicarlos por 10. Así, un tallo de 15 vale 150, un tallo de 18 vale 180, etc.

Las hojas completan la información del tallo. Un tallo de 15 con una hoja de 4 representa una altura de 154 cm; un tallo de 18 con una hoja de 0 representa una altura de 180 cm; etc. El número de casos que representa cada hoja (cada hoja puede representar a más de un caso) se indica en la parte inferior del diagrama (*each leaf*). Así, en la Figura 4.5, *each leaf* = 3 significa que cada hoja representa a 3 casos; en la Figura 4.6, *each leaf* = 1 caso.

Las filas primera y última del diagrama muestran (si los hay) el número de casos con valores *extremos* (*extremes*) y los valores que toman esos casos (entre paréntesis). Así, por ejemplo, en el diagrama de la Figura 4.5 aparecen 2 casos extremos con alturas iguales o menores que 149 cm y 5 casos extremos con alturas iguales o mayores que 192 cm. En el diagrama de la Figura 4.6 aparecen 2 casos extremos con un valor mayor o igual que 168.

Figura 4.5. Diagrama de tallo y hojas de la variable *altura*

```
  2  Extremes (=<149)
  2  14 . 9
  5  15 . 1&
  7  15 . 23
 18  15 . 445555
 29  15 . 6666677777
 35  15 . 888899999999
 52  16 . 000000000111111111
 68  16 . 22222222222333333333333
 79  16 . 444444444444455555555555555
 77  16 . 66666666666667777777777777
100  16 . 888888888888888999999999999999999
112  17 . 0000000000000000000111111111111111111
 88  17 . 22222222222222333333333333333
 84  17 . 4444444444444445555555555555
 74  17 . 666666666677777777777777
 64  17 . 888888888999999999999
 27  18 . 000001111
 29  18 . 2222223333
 16  18 . 44555
 12  18 . 6667
  7  18 . 899
  2  19 . 0
  5  Extremes (>=192)

Stem width: 10,00
Each leaf:  3 case(s)
& denotes fractional leaves.
```

Figura 4.6. Diagrama de tallo y hojas de la variable *tasa de muertes al nacer*

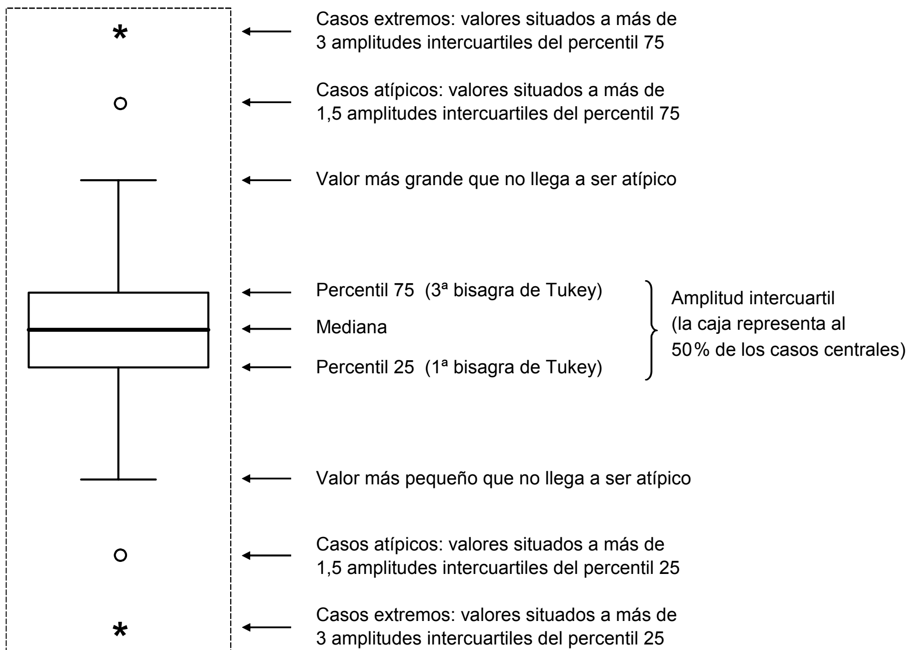
```
27  0 . 445555666666666777778888899
12  1 . 012223467799
16  2 . 0001123555577788
 7  3 . 4567999
 6  4 . 135679
 9  5 . 011222347
 5  6 . 03678
 7  7 . 4556679
 1  8 . 5
 1  9 . 4
 4 10 . 1569
 7 11 . 0022378
 2 12 . 46
 1 13 . 7
 1 Extremes    (>=168)

Stem width: 10,0
Each leaf:  1 case(s)
```

El **diagrama de caja y bigotes** (*box and wiskers plot*) o, simplemente, **diagrama de caja**, es un ingenioso gráfico que permite formarse una idea muy rápida sobre las tres propiedades esenciales de una distribución: centro, dispersión y forma. Incluye la mediana, los percentiles[9] 25 y 75, y una serie de puntos que identifican a los valores que se alejan excesivamente del centro (Tukey, 1977).

La Figura 4.7 muestra los detalles de un diagrama de caja (el diagrama de caja es la parte incluida dentro de la línea de puntos). La mediana identifica el centro de la distribución (debe tenerse en cuenta que el diagrama se representa dentro de un plano cartesiano con los valores de la variable en el eje vertical). La altura de la caja y la longitud de los bigotes permiten valorar el grado de dispersión y de asimetría (los bigotes se extienden hasta lo que podríamos llamar una dispersión *razonable*). Los círculos y los asteriscos, si existen, delatan casos excesivamente alejados del centro.

Figura 4.7. Detalles de un diagrama de caja



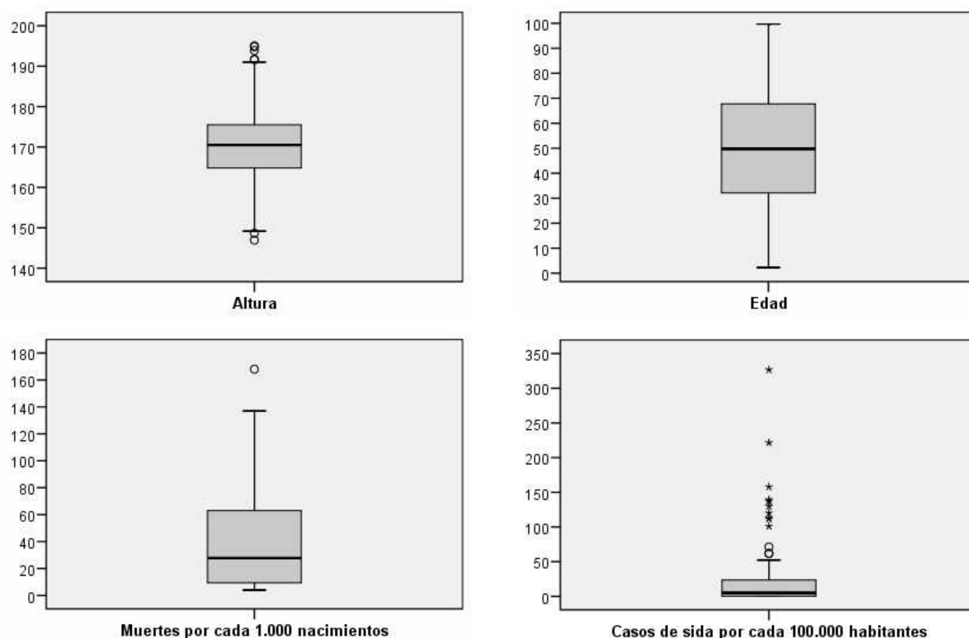
Para ayudar a interpretar los diagramas de caja, la Figura 4.8 muestra las distribuciones ya representadas en los histogramas de la Figura 4.1. En primer lugar intentamos identificar el *centro* de la distribución: puesto que en el eje vertical está representada la escala de la variable, la mediana permite saber que el centro de la distribución de la *altura* se encuentra en torno a 170 cm, que el centro de la distribución de la *edad* se encuentra en torno a 50 años, etc.

[9] En realidad, los diagramas de caja no se construyen con los percentiles 25 y 75, sino con las bisagras primera y tercera de Tukey (valores muy parecidos a los percentiles 25 y 75 pero no siempre idénticos). La primera bisagra es el valor que ocupa la posición intermedia entre la mediana y el valor más pequeño; la tercera bisagra es el valor que ocupa la posición intermedia entre la mediana y el valor más grande.

En segundo lugar, para valorar el grado de *dispersión* hay que interpretar correctamente la altura de la caja y la longitud de los bigotes; para ello, debe tenerse en cuenta que los diagramas están dibujados ocupando toda la altura de su correspondiente recuadro; por tanto, la altura total del gráfico es irrelevante; donde hay que fijarse es en la longitud relativa de los bigotes, es decir, en la longitud de los bigotes comparada con la altura de la caja; si el bigote se extiende todo lo que puede extenderse (1,5 amplitudes intercuartiles; es decir, 1,5 veces la altura de la caja), es que hay casos que llegan hasta ahí; si el bigote no alcanza su longitud máxima, es que no existen casos en esa dirección. Sabemos, por ejemplo, que la dispersión de la distribución de la *altura* es mayor que la dispersión de la distribución de la *edad* (ver histogramas de la Figura 4.1). Esta diferencia en la dispersión está reflejada, por un lado, en la longitud de los bigotes (los bigotes alcanzan su máxima extensión en la distribución de la *altura*, pero no en la de la *edad*) y, por otro, en la presencia de casos atípicos y extremos (círculos y asteriscos por encima y por debajo de los bigotes) en la distribución de la *altura*.

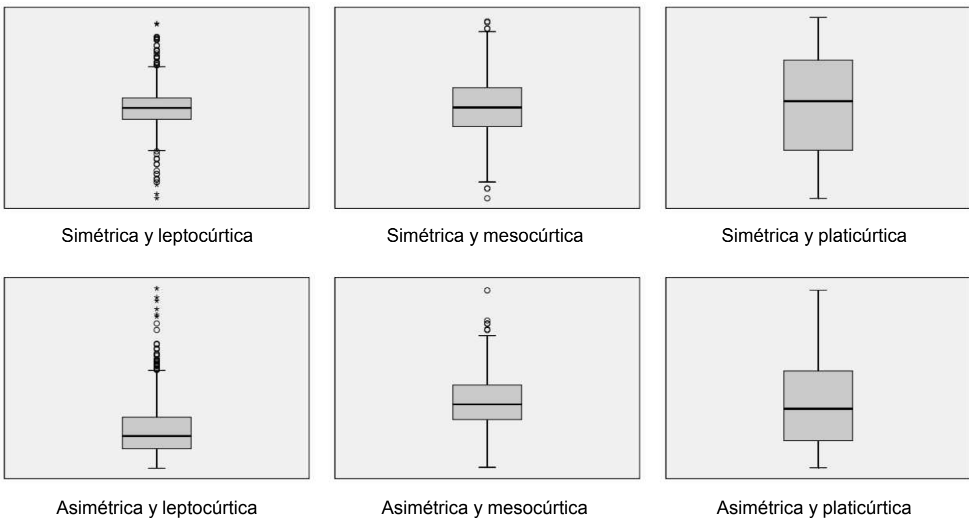
Por último, la longitud de los dos bigotes de cada diagrama y la presencia o no de casos atípicos y extremos están informando de la *forma de la distribución*. En la distribución de la *altura* se observan bigotes de igual longitud y un número aproximadamente igual de casos atípicos y extremos en ambas colas de la distribución; esto indica que se trata de una distribución aproximadamente simétrica. Lo mismo vale decir de la distribución de la *edad* (bigotes de igual longitud y mismo número –ninguno– de casos atípicos o extremos). Esta pauta es bastante distinta de la que ofrecen los otros dos diagramas: en ambos ocurre que el bigote superior es mucho mayor que el inferior, lo cual indica que las distribuciones son claramente asimétricas; pero, además, el número de casos atípicos y extremos es mucho mayor en la distribución de la *tasa de sida*, lo que indica que esta distribución es mucho más asimétrica que la de la *tasa de muertes al nacer*.

Figura 4.8. Diagramas de caja de las variables *altura*, *edad*, *muertes al nacer* y *tasa de sida*



En los diagramas de caja de la Figura 4.9 están representadas las mismas distribuciones que en los histogramas de la Figura 4.3. Los diagramas de la mitad superior representan tres distribuciones simétricas (bigotes de igual longitud) con distinto grado de curtosis: el diagrama de la izquierda muestra una distribución leptocúrtica (muchos casos fuera de los bigotes), el del centro una distribución mesocúrtica (sólo unos pocos casos fuera de los bigotes) y el de la derecha una distribución platicúrtica (no hay casos fuera de los bigotes). Los diagramas de la mitad inferior representan tres distribuciones asimétricas (bigotes de longitud desigual) con distinto grado de curtosis: el diagrama de la izquierda muestra una distribución leptocúrtica (casos atípicos y extremos sólo en la parte superior), el del centro una distribución mesocúrtica (casos atípicos sólo en la parte superior) y el de la derecha una distribución platicúrtica (ni casos atípicos ni casos extremos).

Figura 4.9. Diagramas de caja: distribuciones con distinto grado de asimetría y curtosis



Índices de asimetría y curtosis

Aunque la inspección de un gráfico (histograma, diagrama de caja) ya permite formarse una primera idea sobre el grado de asimetría y curtosis de una distribución, utilizar estadísticos permite valorar ambas características con mayor precisión.

Para cuantificar el grado de asimetría y de curtosis de una distribución se han propuesto múltiples índices (puede encontrarse una buena recopilación en Solanas, Salafranca, Fauquet y Núñez, 2005, págs. 321-327 y 353-358). De todos ellos, en este apartado únicamente se incluyen los dos que ofrece el SPSS. Ambos se basan en el concepto de *momento respecto a la media* (importante concepto estadístico relacionado con las desviaciones de la media, es decir, con las puntuaciones diferenciales o de desviación). El momento de orden r respecto a la media se define de la siguiente manera:

$$M_r = \sum (Y_i - \bar{Y})^r = \sum y_i^r \qquad [4.17]$$

El momento de orden 1 vale cero (recordemos que las desviaciones de la media suman cero; ver ecuación [4.3]). El momento de orden 2 es el numerador de la fórmula de la varianza. El momento de orden 3 constituye la base de un **índice de asimetría** propuesto inicialmente por Fisher y que en el SPSS se calcula mediante

$$g_1 = \frac{n\,M_3}{(n-1)(n-2)\,S_Y^3} \qquad [4.18]$$

(Bliss, 1967, pág. 144). Y el momento de orden 4 constituye la base de un **índice de curtosis** también propuesto originalmente por Fisher y que en el SPSS se calcula mediante

$$g_2 = \frac{n(n+1)M_4 - 3(n-1)M_2^2}{(n-1)(n-2)(n-3)\,S_Y^4} \qquad [4.19]$$

Ambos índices valen cero cuando la distribución es simétrica y mesocúrtica. Índices de asimetría (g_1) mayores que cero indican asimetría positiva; índices menores que cero indican asimetría negativa. Índices de curtosis (g_2) mayores que cero indican leptocurtosis; índices menores que cero indican platicurtosis. Calculados en las distribuciones representadas en las Figuras 4.3 y 4.9 se obtienen los resultados que muestra la Tabla 4.7.

Cuando las distribuciones son aproximadamente simétricas (las tres primeras), los índices de asimetría toman valores próximos a cero (en el ejemplo, entre −0,104 y 0,088); cuando las distribuciones son asimétricas (las tres últimas), toman valores más alejados de cero (en el ejemplo, entre 0,437 y 1,519).

Cuando las distribuciones son aproximadamente mesocúrticas (la segunda y la quinta), los índices de curtosis toman valores próximos a cero (en el ejemplo, −0,087 y 0,166); cuando las distribuciones son leptocúrticas (la primera y la cuarta), toman valores positivos alejados de cero (en el ejemplo, 2,570 y 3,436); cuando las distribuciones son platicúrticas (la tercera y la sexta), toman valores negativos alejados de cero (en el ejemplo, −1,175 y −0,637).

Tabla 4.7. Índices de asimetría (g_1) y curtosis (g_2) calculados en las distribuciones de las Figuras 4.3 y 4.9

	Simétrica leptocúrtica	*Simétrica mesocúrtica*	*Simétrica platicúrtica*	*Asimétrica leptocúrtica*	*Asimétrica mesocúrtica*	*Asimétrica platicúrtica*
g_1	0,088	0,05	−0,104	1,519	0,437	0,543
g_2	2,57	−0,087	−1,175	3,436	0,166	−0,637

Expresiones del tipo “un valor próximo a cero” o “un valor más alejado de cero” ayudan poco a tomar una decisión sobre el grado de asimetría o curtosis de una distribución. Para resolver este problema puede recurrirse a una sencilla estrategia que, aunque todavía no puede explicarse con detalle (se basa en conceptos inferenciales que se explican más adelante), permite tomar decisiones muy rápidas. La estrategia consiste en dividir el índice de asimetría (o el de curtosis) entre su desviación típica. Si el resultado se encuentra entre −2 y 2 puede asumirse que la distribución es simétrica (o mesocúrtica); si es mayor que 2, puede afirmarse que la distribución es asimétrica positiva (o leptocúrtica); y si es menor que −2, puede afirmarse que la distribución es asimétrica negativa (o platicúrtica). Las desviaciones típicas de los índices

de asimetría y curtosis (las cuales, simplemente por estar referidas a *estadísticos* cambian el nombre de desviación típica por el de *error típico*), pueden obtenerse mediante:

$$S_{g_1} = \sqrt{\frac{6n(n-1)}{(n+1)(n-2)(n+3)}} \quad \text{y} \quad S_{g_2} = \sqrt{\frac{4(n^2-1)\,S_{g_1}^2}{(n-3)(n+5)}} \qquad [4.20]$$

Estas ecuaciones ofrecen un valor de 0,077 para el error típico del índice de asimetría y un valor 0,155 para el del índice de curtosis (son los mismos para las seis distribuciones porque todas ellas tienen 1.000 casos). En la distribución simétrica y mesocúrtica de la Figura 4.9 (la segunda), el cociente entre el índice de asimetría y su error típico vale 0,050/0,077 = 0,65; y el cociente entre el índice de curtosis y su error típico vale −0,087/0,155 = −0,56. Puesto que ambos valores se encuentran entre −2 y 2, puede asumirse que se trata de una distribución simétrica y mesocúrtica. Sin embargo, en la distribución asimétrica y leptocúrtica (la cuarta), el cociente entre el índice de asimetría y su error típico vale 1,519/0,077 = 19,72; y el cociente entre el índice de curtosis y su error típico vale 3,436/0,155 = 22,17. Ahora ambos valores son mayores que 2 (mucho mayores que 2) y esto permite afirmar que se trata de una distribución con una fuerte asimetría positiva y mucha leptocurtosis.

Análisis descriptivo de variables cuantitativas con SPSS

Aunque la mayoría de los procedimientos SPSS incluyen opciones que permiten obtener los estadísticos descriptivos más comúnmente utilizados, hay algunos procedimientos que están específicamente diseñados para obtener este tipo de estadísticos. No existe, sin embargo, un único procedimiento que incluya todos los estadísticos estudiados en este capítulo.

El procedimiento **Frecuencias**, además de tablas de frecuencias y gráficos de barras y sectores (herramientas ya estudiadas en el capítulo anterior para describir variables categóricas), también permite obtener algunos de los estadísticos estudiados en este capítulo (algunos estadísticos de tendencia central como la media y la mediana; algunos de dispersión como la amplitud total, la varianza y la desviación típica; algunos estadísticos y gráficos sobre la forma de la distribución como el índice de asimetría, el de curtosis y los histogramas; y todos los estadísticos de posición o cuantiles), pero no permite obtener otros muchos.

Para poder obtener los estadísticos estudiados en este capítulo vamos a revisar los procedimientos **Explorar** y **Razón** centrándonos en la información descriptiva que ofrece cada uno de ellos. El procedimiento **Explorar** incluye algunas opciones no disponibles en otros procedimientos. En relación con las herramientas descriptivas, a los estadísticos y gráficos que ofrece el procedimiento **Frecuencias** añade la media recortada, los estimadores robustos centrales o estimadores *M* (Andrews, Hampel, Huber y Tukey), la amplitud intercuartil, el diagrama de tallo y hojas, y el diagrama de caja. Además, permite obtener todos estos estadísticos y gráficos para los subgrupos definidos por una variable categórica.

El procedimiento **Razón** ofrece dos estadísticos de dispersión que no se encuentran en los procedimientos **Frecuencias** y **Explorar**: el coeficiente de variación centrado en la media y el centrado en la mediana. Aunque el procedimiento está diseñado para analizar el cociente entre dos variables, colocando en el denominador una variable cuyos valores sean "unos" puede utilizarse para analizar variables individuales. También permite obtener resultados para los subgrupos definidos por una variable categórica.

Ejemplo. Tendencia central, dispersión y forma de la distribución con SPSS

Este ejemplo muestra cómo obtener con el SPSS los estadísticos y gráficos estudiados en este capítulo. El análisis se basa en el archivo *Datos de empleados* ya utilizado en el capítulo anterior (puede encontrarse en la misma carpeta en la que está instalado el SPSS).

La mayor parte de los estadísticos estudiados en este capítulo pueden obtenerse con el procedimiento **Explorar**. Para los estadísticos no incluidos en **Explorar** utilizaremos el procedimiento **Razón**. El procedimiento **Frecuencias** ya lo hemos aplicado en el capítulo anterior y, en lo relativo a la descripción de variables cuantitativas, no añade nada al procedimiento **Explorar**. Para aplicar el procedimiento **Explorar**:

➤ Seleccionar la opción **Estadísticos descriptivos > Explorar** del menú **Analizar** para acceder al cuadro de diálogo **Explorar** y trasladar la variable *salario* (salario actual) a la lista **Dependientes** y la variable *sexo* (sexo del empleado) a la lista **Factores** (la lista **Dependientes** únicamente admite variables con formato numérico; la lista **Factores** admite variables con formato numérico y de cadena corta).

➤ Pulsar el botón **Estadísticos** para acceder al subcuadro de diálogo **Explorar: Estadísticos** y marcar las opciones **Descriptivos** (está marcada por defecto), **Estimadores robustos centrales**, **Valores atípicos** y **Percentiles**; pulsar el botón **Continuar** para volver al cuadro de diálogo principal.

➤ Pulsar el botón **Gráficos** para acceder al subcuadro de diálogo **Explorar: Gráficos** y seleccionar la opción **Dependientes juntas** del recuadro **Diagramas de caja** y la opción **Histograma** del recuadro **Descriptivos**; pulsar el botón **Continuar** para volver al cuadro de diálogo principal.

Aceptando estas selecciones el *Visor* ofrece los resultados que muestran las Tablas 4.8 a la 4.13 y las Figuras 4.10 y 4.11. Ya sabemos que, para describir correctamente la variable *salario actual* (al igual que cualquier otra variable cuantitativa), debemos prestar atención a tres propiedades básicas: centro, dispersión y forma de la distribución.

Los estadísticos que permiten identificar el ***centro*** de la distribución están repartidos entre las Tablas 4.8 y 4.9. Comenzando con la distribución de los hombres, se observa que la media es el estadístico que mayor salario asigna al centro de la distribución (41.441,78), la media recortada al 5% asigna un valor algo inferior (39.445,87) y la mediana un valor sensiblemente inferior (32.850). Los estimadores *M* de la Tabla 4.9 oscilan entre 31.722,27 y 34.820,15, lo cual está indicando que el valor de la mediana permite identificar el centro de la distribución del salario mejor de lo que lo hacen la media y la media recortada.

Según los estadísticos de ***dispersión*** (Tabla 4.8), las distancias a la media valen, en promedio, 19.500 dólares (*desv. típ.* = 19.499,21). Quizá este valor no permita formarse una idea precisa acerca del grado de dispersión existente, pero en esto puede ayudar la amplitud total, que toma un valor en torno a 115.000 dólares (*rango* = 115.350,00). Si se tiene en cuenta que el 50% de los casos centrales se encuentra en un rango de unos 22.500 dólares (*amplitud intercuartil* = 22.675,00), entonces una amplitud de 115.000 dólares está delatando la presencia de mucha dispersión (más tarde, revisando los gráficos y los casos con valores más extremos habrá que matizar este resultado). La varianza, finalmente, toma un valor desprovisto por completo de utilidad descriptiva; en una distribución cuyo centro se encuentra situado no lejos de 35.000 dólares, un valor de 380.219.336,30 no tiene ningún significado.

En relación con la ***forma de la distribución***, la Tabla 4.8 ofrece los índices de asimetría y curtosis con sus respectivos errores típicos. El índice de asimetría (*asimetría* = 1,64) dividido por su error típico (*error típ.* = 0,15) vale 10,93, lo cual significa que existe un elevado grado de asimetría positiva (pues 10,9 > 2); esto es algo que ya podríamos haber anticipado al conocer que la media tomaba un valor sensiblemente mayor que la mediana. Y el índice de curtosis (*curtosis* = 2,78) dividido entre su error típico (*error típ.* = 0,30) vale 9,3, lo cual significa que nos encontramos ante una distribución leptocúrtica (9,3 > 2), es decir, una distribución en la que una de sus colas (ya sabemos que la distribución es asimétrica) concentra más casos de los que concentra la correspondiente curva normal.

Tabla 4.8. Estadísticos descriptivos del procedimiento *Explorar*

Salario actual

Sexo		Estadístico	Error típ.
Hombre	Media	41.441,78	
	Media recortada al 5%	39.445,87	
	Mediana	32.850,00	
	Varianza	380.219.336,30	
	Desv. típ.	19.499,21	
	Mínimo	19650,00	
	Máximo	135000,0	
	Rango	115.350,00	
	Amplitud intercuartil	22.675,00	
	Asimetría	1,64	,15
	Curtosis	2,78	,30
Mujer	Media	26.031,92	
	Media recortada al 5%	25.248,30	
	Mediana	24.300,00	
	Varianza	57.123.688,27	
	Desv. típ.	7.558,02	
	Mínimo	15750,00	
	Máximo	58125,00	
	Rango	42.375,00	
	Amplitud intercuartil	7.012,50	
	Asimetría	1,86	,17
	Curtosis	4,64	,33

Tabla 4.9. Estimadores robustos centrales (estimadores *M*)

Salario actual

Sexo	Estimador-M de Huber[a]	Biponderado de Tukey[b]	Estimador-M de Hampel[c]	Onda de Andrews[d]
Hombre	34.820,15	31.779,76	34.020,57	31.732,27
Mujer	24.606,10	24.015,98	24.419,25	24.005,82

a. La constante de ponderación es 1,339.
b. La constante de ponderación es 4,685.
c. Las constantes de ponderación son 1,700, 3,400 y 8,500.
d. La constante de ponderación es 1,340*pi.

En la distribución de las mujeres se observa una pauta muy similar a la que se observa en la de los de hombres, pero menos pronunciada. La media sigue siendo el estadístico que mayor valor asigna al centro de la distribución (26.031,92), pero su diferencia con la media recortada

al 5% (25.248,30) y con la mediana (24.300,00) es mucho menor que la encontrada en la distribución de los hombres. También la dispersión es ahora menor que en el caso de los hombres (*desv. típ.* = 7.558,02; *rango* = 42.375,00), pero sigue tratándose de una distribución muy asimétrica (1,86/0,17 = 10,94) y leptocúrtica (4,64/0,33 = 14,1).

De todo lo anterior cabe resumir que la variable *salario* en la distribución de los hombres tiene un centro situado en torno a 33.000 dólares, con mucha dispersión y fuerte asimetría; y, en la de las mujeres, un centro situado en torno a 24.300 dólares, con menor dispersión pero también con una fuerte asimetría. Los gráficos que veremos a continuación (Figuras 4.10 y 4.11) permitirán completar esta primera impresión.

La Tabla 4.10 ofrece algunos percentiles y los cuartos o bisagras de Tukey (si se desea obtener percentiles distintos de los que ofrece el procedimiento **Explorar**, puede utilizarse el procedimiento **Frecuencias**). Recordemos que los percentiles sirven, por un lado, para ubicar a cada sujeto en la posición relativa que ocupa en su grupo de referencia (a modo de baremos o tablas de clasificación) y, por otro, para comparar entre sí puntuaciones individuales de distintos grupos o variables. ¿Qué puede decirse de un empleado cuyo salario es de 24.000 dolares? Pues, si es un hombre, se trata de un empleado que se encuentra por debajo del percentil 10 (P_{10}=25.500 dólares); por tanto, más del 90% de los empleados de su mismo sexo tienen salarios mayores que el suyo. Sin embargo, si es una mujer, se trata de una empleada situada aproximadamente en el centro de su grupo (P_{50}=24.300 dólares). Las bisagras o cuartos de Tukey coinciden con los cuartiles: la segunda bisagra siempre coincide con la mediana; las bisagras primera y tercera coinciden con el primer y tercer cuartil, pero pueden diferir ligeramente (es muy raro encontrar informes en los que se utilicen las bisagras de Tukey en lugar de los cuartiles).

Tabla 4.10. Percentiles y bisagras de Tukey

Salario actual

		Percentiles						
Sexo		5	10	25	50	75	90	95
Hombre	Promedio pond.	23.212,50	25.500,00	28.050,00	32.850,00	50.725,00	69.325,00	81.312,50
	Bisagras de Tukey			28.050,00	32.850,00	50.550,00		
Mujer	Promedio pond.	16.950,00	18.660,00	21.487,50	24.300,00	28.500,00	34.890,00	40.912,50
	Bisagras de Tukey			21.525,00	24.300,00	28.500,00		

La Tabla 4.11 (la tabla original se ha pivotado para adaptarla a las dimensiones de la página) ofrece un listado de los valores más pequeños y más grandes (5 de cada tipo) de cada grupo. Aunque el SPSS los llama *atípicos*, en realidad sólo se trata de los 5 valores más pequeños y los 5 más grandes; lo cual no significa que sean atípicos en el sentido de anómalos o muy alejados del centro; en una distribución simétrica y platicúrtica, podrían ser valores cercanos al centro. Este listado sirve, en primer lugar, para detectar posibles errores en los datos: valores excesivamente pequeños o excesivamente grandes como consecuencia, por ejemplo, de haber puesto un cero de más o un cero de menos en el salario, aparecerán en este listado. También sirve este listado para comprobar si existe algún valor que, aun no siendo un error, se encuentra excesivamente alejado de los demás. Entre los valores más grandes del grupo de hombres se observa un valor sensiblemente mayor que los restantes (el caso 29 tiene un salario de 135.000 dólares; el segundo valor mayor es de 110.625); en el grupo de mujeres no se observa este salto. Tampoco se observan saltos entre los 5 valores más pequeños (ni en el grupo de hombres ni en el de mujeres). Por supuesto, eliminar el caso 29 del análisis podría

mejorar nuestra caracterización del salario de los hombres, pues tanto la dispersión como la asimetría quedarían reducidas.

Tabla 4.11. Valores atípicos

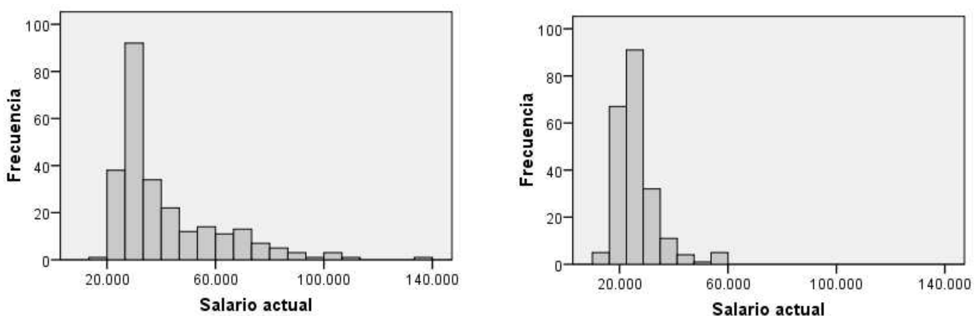
Salario actual

Sexo			1	2	3	4	5
Hombre	Mayores	Número del caso	29	32	18	343	446
		Valor	135000,0	110625,0	103750,0	103500,0	100000,0
	Menores	Número del caso	192	372	258	22	65
		Valor	19650,00	21300,00	21300,00	21750,00	21900,00
Mujer	Mayores	Número del caso	371	348	468	240	72
		Valor	58125,00	56750,00	55750,00	54375,00	54000,00
	Menores	Número del caso	378	338	411	224	90
		Valor	15750,00	15900,00	16200,00	16200,00	16200,00

Los histogramas de la Figura 4.10 no hacen otra cosa que confirmar visualmente lo que ya nos han dicho los números. Ya sabemos que tanto el centro como la dispersión del salario son mayores en la distribución de los hombres que en la de las mujeres; y también sabemos que ambas distribuciones son muy asimétricas y leptocúrticas. Pues bien, los valores del eje horizontal del histograma indican que tanto el centro como la dispersión de las distribuciones son mayores en el caso de los hombres que en el de mujeres. En relación con el centro de la distribución, el punto de equilibrio del histograma de los hombres se encuentra a medio camino entre 20.000 y 60.000 dólares, mientras que en el caso de las mujeres se encuentra mucho más cerca de 20.000 dólares. Por lo que se refiere a la dispersión, en la distribución de los hombres se observan muchos casos por encima de 60.000 dólares, con algunos llegando a sobrepasar los 100.000, mientras que en la de las mujeres no hay casos por encima de los 60.000. Por último, los histogramas también están indicando que en ambos casos se trata de distribuciones muy asimétricas (asimetría positiva) y leptocúrticas[10].

Los diagramas de tallo y hojas ofrecen la misma información que el histograma, pero con más detalle (no se incluyen aquí).

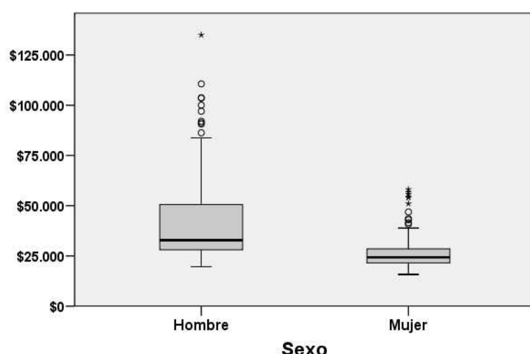
Figura 4.10. Histogramas de *salario actual* para *hombres* (izquierda) y *mujeres* (derecha)



[10] La valoración del grado de asimetría y curtosis de una distribución a partir de su histograma es más fácil si al histograma se le superpone una curva normal. Esto puede hacerse entrando en el *Editor de gráficos* (pinchando dos veces sobre el gráfico que se desea editar) y seleccionando la opción **Mostrar curva de distribución** del menú **Elementos**.

Finalmente, los diagramas de caja de la Figura 4.11 corroboran lo que venimos observando. En primer lugar, las medianas (líneas horizontales que atraviesan las cajas) están indicando que el centro de la distribución de los hombres es mayor que el de la distribución de las mujeres. En segundo lugar, la altura de las cajas y la longitud de los bigotes está indicando que la distribución de los hombres es mucho más dispersa que la de las mujeres (a las distribuciones con dispersión similar les corresponden cajas de altura similar y bigotes de longitud similar). Por último, la presencia de casos atípicos y extremos en la parte alta de la distribución (valores por encima del bigote superior) y los bigotes cortos en la parte baja de la distribución, están indicando que se trata, en ambos casos, de distribuciones con asimetría positiva.

Figura 4.11. Diagramas de caja de *salario actual* para *hombres* y *mujeres*



El procedimiento **Explorar** no incluye las medidas de dispersión relativa estudiadas en este capítulo: el coeficiente de variación centrado en la media y el centrado en la mediana. Para obtener estos estadísticos es necesario utilizar el procedimiento **Razón**. Antes de poder utilizar este procedimiento para analizar una sola variable hay que crear una variable cuyos valores sean todo "unos"[11]. Hecho esto:

➤ Seleccionar la opción **Estadísticos descriptivos > Razón** del menú **Analizar** para acceder al cuadro de diálogo **Estadísticos de la razón** y trasladar la variable *salario* (salario actual) al cuadro **Numerador** y la variable cuyos valores son todo "unos" (la variable que hemos creado nosotros) al cuadro **Denominador** (estos cuadros de selección de variables únicamente admiten variables con formato numérico). Trasladar al cuadro **Variable de agrupación** la variable *sexo* (sexo del empleado).

➤ Pulsar el botón **Estadísticos** para acceder al subcuadro de diálogo **Estadísticos de la razón: Estadísticos** y, de todas las opciones disponibles, marcar únicamente **CDV centrado en la media** y **CDV centrado en la mediana**; pulsar el botón **Continuar** para volver al cuadro de diálogo principal.

Aceptando estas selecciones el *Visor* ofrece los resultados que muestra la Tabla 4.12. El coeficiente de variación centrado en la media indica que el salario de los hombres es, en términos relativos, más disperso (47,1%) que el de las mujeres (29,0%). El coeficiente de variación

[11] Esta variable puede crearse fácilmente mediante la opción **Calcular** del menú **Transformar** utilizando como expresión numérica la constante 1.

centrado en la mediana indica exactamente lo mismo. En esta comparación se está teniendo en cuenta el hecho de que el salario medio de los hombres es mayor que el de las mujeres (cosa que no se había hecho con el resto de estadísticos de dispersión). Puesto que el salario medio de los hombres es mayor que el de las mujeres, es razonable esperar que la dispersión de los salarios sea mayor en el grupo de hombres que en el de mujeres. Los coeficientes de variación indican que esa diferencia en las dispersiones se mantiene incluso cuando se anula el efecto de la diferencia entre las medias.

Obsérvese que, como la distribución del salario es asimétrica positiva, el coeficiente centrado en la media es menor que el centrado en la mediana.

Tabla 4.12. Media de las desviaciones absolutas y coeficientes de variación

	Coeficiente de variación	
Grupo	Centrado en la media	Centrado en la mediana
Hombre	47,1%	64,9%
Mujer	29,0%	31,9%
Global	49,6%	62,2%

El SPSS no calcula directamente algunos de los estadísticos incluidos en este capítulo (no tiene cuadros de diálogo con opciones para calcularlos). En concreto, no calcula la media winsorizada, la trimedia, la media de las desviaciones y la mediana de las desviaciones. No obstante, en el caso de que interese calcular estos estadísticos, es posible hacerlo utilizando la sintaxis del programa. En la página *web* del manual se puede encontrar un archivo con la sintaxis SPSS necesaria para obtener estos cuatro estadísticos.

Análisis descriptivo y exploratorio

A lo largo de todo este capítulo hemos intentado enfatizar la idea de que describir correctamente una variable cuantitativa requiere prestar atención a tres propiedades básicas de su distribución (centro, dispersión y forma) y hemos presentado las herramientas descriptivas más utilizadas para abordar el estudio de esas tres características. Esto es, sin duda, la esencia del análisis descriptivo.

Pero las herramientas descriptivas, además de ofrecer una caracterización apropiada de los datos, poseen una utilidad añadida. Independientemente de la complejidad de los datos disponibles y del procedimiento estadístico que finalmente se tenga intención de aplicar, una exploración minuciosa de los datos con herramientas descriptivas, previa al inicio de cualquier otro tipo de análisis, posee importantes ventajas que un analista de datos no puede pasar por alto (ver Behrens, 1997). Una exploración descriptiva de los datos permite identificar, entre otras cosas, posibles errores (datos mal introducidos, respuestas mal codificadas, etc.), valores atípicos (valores que se alejan demasiado del resto), pautas extrañas en los datos (valores que se repiten demasiado o que no aparecen nunca, etc.), variabilidad no esperada (demasiada concentración en torno a determinado valor, demasiados casos en una de las dos colas de la distribución), etc.

Las influyentes obras de Tukey (1977) y Hoaglin, Mosteller y Tukey (1983) han llamado la atención de los investigadores sobre los importantes beneficios de esta exploración inicial de los datos. Por esta razón, en este capítulo sobre análisis descriptivo de variables cuantitativas, hemos recogido, además de los estadísticos clásicos diseñados para describir el centro, la dispersión y la forma de la distribución, algunas de las herramientas *exploratorias* (estadísticos resistentes, diagramas de tallo y hojas, de caja, etc.) que, además de ayudar a describir los datos, permiten efectuar una exploración de los mismos con el objetivo, no sólo de describirlos, sino de detectar posibles anomalías.

Apéndice 4

Reglas del sumatorio

Entender muchas de las fórmulas que se utilizan en estadística requiere estar familiarizado con el símbolo del sumatorio (Σ). Este apartado describe su significado y las reglas básicas a las que se ajusta.

Recordemos que las variables se representan con letras latinas mayúsculas (X, Y, Z, etc.) y los valores concretos que toman con letras latinas mayúsculas acompañadas de un subíndice (X_i, Y_i, Z_i, etc.). El subíndice no tiene nada que ver con los valores concretos que toma la variable sino con la posición que ocupan esos valores: Y_1 se refiere al valor que toma el primer elemento; Y_2 al valor que toma el segundo elemento; Y_n al valor que toma el último elemento. Así, si la variable Y toma los valores 3, 7, 9, 12 y 15, entonces: $Y_1 = 3$, $Y_2 = 7$, ..., $Y_5 = Y_n = 15$. En este caso, el subíndice i toma los valores 1, 2, 3, 4 y 5, lo cual se representa mediante: $i = 1, 2, ..., 5$. Para sumar esos 5 valores se utiliza la expresión

$$\sum_{i=1}^{5} Y_i$$

Por tanto, esta expresión (que se lee "suma o sumatorio de Y") significa:

$$\sum_{i=1}^{5} Y_i \;=\; Y_1 + Y_2 + Y_3 + Y_4 + Y_5 \;=\; 3 + 7 + 9 + 12 + 15 \;=\; 46$$

Del mismo modo que el signo de la suma (+) indica que hay que sumar *dos valores*, el símbolo del sumatorio (Σ) indica que hay que sumar un *determinado número de valores*. Si una variable toma valores de 1 a n (es decir, $i = 1, 2, ..., n$), la suma de los n valores se expresa de la siguiente manera:

$$\sum_{i=1}^{n} Y_i \;=\; Y_1 + Y_2 + Y_3 + \cdots + Y_n$$

El subíndice y el superíndice del símbolo del sumatorio indican que hay que sumar comenzando con el primer valor ($i = 1$) y terminando con el último (n). Cuando se quiere expresar la suma de todos los elementos (desde $i = 1$ hasta $i = n$) pueden eliminarse el subíndice y el superíndice del sumatorio:

$$\sum Y_i \;=\; \sum_{i=1}^{n} Y_i$$

Conocer una sencillas reglas ayuda a simplificar bastante el trabajo con el símbolo del sumatorio. La primera de ellas puede formularse así: *el sumatorio de una constante es igual a n veces esa constante*:

$$\sum c \;=\; nc$$

Otra regla útil es la que afirma que *el sumatorio de una variable multiplicada por una constante es igual al producto de la constante por el sumatorio de la variable*:

$$\sum cY_i = c\sum Y_i$$

Además, *el sumatorio de una suma es la suma de los correspondientes sumatorios*:

$$\sum (X_i + Y_i) = \sum X_i + \sum Y_i$$

Esto no ocurre con el sumatorio de un producto; en general, el sumatorio de un producto es distinto del producto de los correspondientes sumatorios: $\Sigma(X_i Y_i) \neq \Sigma X_i \Sigma Y_i$. Combinando las tres reglas anteriores se deduce, por ejemplo, que

$$\sum (cY_i + k) = c\sum Y_i + nk$$

Y también de las reglas anteriores se deduce que

$$\sum (Y_i + k)^2 = \sum (Y_i^2 + k^2 + 2kY_i) = \sum Y_i^2 + nk^2 + 2k\sum Y_i$$

En general, la suma de los valores elevados al cuadrado es distinta del cuadrado de la suma de los valores; es decir, $\Sigma Y_i^2 \neq (\Sigma Y_i)^2$.

Cuando una variable se mide en varios grupos es necesario utilizar dos subíndices para poder identificar cada valor de la variable. La Tabla 4.13 puede ayudar a entender esto. Identificar la posición de un valor dentro de un grupo requiere utilizar, como hasta ahora, el subíndice i. Identificar a qué grupo pertenece ese valor requiere utilizar un subíndice adicional; suele utilizarse el subíndice j. Así, la puntuación Y_{ij} se refiere a la puntuación que ocupa la posición i en el grupo j. Llamando J al número de grupos, el subíndice j tomará valores desde 1 hasta J; es decir, $j = 1, 2, ..., J$. En el ejemplo de la Tabla 4.13, $j = 1, 2, 3$; $J = 3$.

Tabla 4.13. Edades de tres grupos de sujetos

Grupos	*Edades*												n_j
$j = 1$	19	20	22	23	24	25	26	27	28				9
$j = 2$	18	19	20	22	24	26	29	31	35	45	61		11
$j = 3$	18	19	20	22	24	26	29	33	35	36	41	43	12
Posición	$i=1$	$i=2$	$i=3$	$i=4$	$i=5$	$i=6$	$i=7$	$i=8$	$i=9$	$i=10$	$i=11$	$i=12$	

Cada grupo puede tener el mismo o distinto tamaño, por tanto, el subíndice i, en el primer grupo, tomará valores desde 1 hasta n_1; en el segundo grupo, desde 1 hasta n_2; en el j-ésimo grupo (esta es la forma de referirnos a uno cualquiera de los grupos), desde 1 hasta n_j; en el último grupo, desde 1 hasta n_J. De acuerdo con esta notación, la suma de las puntuaciones del primer grupo vendrá representada por

$$\sum_{i=1}^{n_1} Y_{i1}$$

El subíndice i indica que hay que sumar todas las puntuaciones desde $i = 1$ hasta n_1; el subíndice j indica que esa suma hay que hacerla únicamente en el primer grupo ($j = 1$). La suma de las puntuaciones del segundo grupo vendrá dada por

$$\sum_{i=1}^{n_2} Y_{i2}$$

El subíndice i indica que hay que sumar todas las puntuaciones desde $i = 1$ hasta n_2; el subíndice j indica que esa suma hay que hacerla únicamente en el segundo grupo ($j = 2$). El sumatorio de las puntuaciones de los restantes grupos se representa siguiendo la misma lógica. Así, en el ejemplo de la Tabla 4.13:

$$\sum_{i=1}^{n_1} Y_{i1} = 19+20+22+23+24+25+26+27+28 = 214$$

$$\sum_{i=1}^{n_2} Y_{i2} = 18+19+20+22+24+26+29+31+35+45+61 = 330$$

$$\sum_{i=1}^{n_3} Y_{i3} = 18+19+20+22+24+26+29+33+35+36+41+43 = 346$$

Para representar la suma de las puntuaciones de un grupo cualquiera se utiliza la expresión general

$$\sum_{i=1}^{n_j} Y_{ij}$$

La suma de todas las puntuaciones (las de todos los grupos) requiere combinar dos símbolos de sumatorio:

$$\sum_{i=1}^{n_j} \sum_{j=1}^{J} Y_{ij} = \sum_{j=1}^{J} \sum_{i=1}^{n_j} Y_{ij}$$

El sumatorio con el subíndice i indica que deben sumarse todos los valores i de cada grupo; el sumatorio con el subíndice j indica que deben sumarse los resultados de cada grupo. En el ejemplo de la Tabla 4.13, la suma de todas las puntuaciones de la tabla se expresa mediante:

$$\sum_{j=1}^{J} \sum_{i=1}^{n_j} Y_{ij} = 214+330+346 = 890$$

Métodos para el cálculo de cuantiles

El SPSS incluye cinco métodos para el cálculo de cuantiles. El método *haverage* es el que se aplica por defecto (ver ecuación [4.1]). El método *waverage* es idéntico en todo al *haverage* excepto en un detalle: $i = kn/100$. El método *round* asigna al cuantil buscado el valor que ocupa la posición entera más próxima a $i = kn/100$. El método *empirical* asigna el valor que ocupa la posición $i = kn/100$ cuando i es un número entero y el que ocupa la posición siguiente a la parte entera de i cuando $i = kn/100$ es un número decimal. El método *aempirical* asigna la media de Y_i e Y_{i+1} cuando $i = kn/100$ es un número entero y el que ocupa la posición siguiente a la parte entera de i cuando $i = kn/100$ es un número decimal.

Para aplicar estos distintos métodos de cálculo de cuantiles es necesario utilizar la sintaxis (desde los cuadros de diálogo sólo es posible aplicar el método *haverage*). Ejecutando las siguientes sentencias es posible obtener los percentiles 5, 10, 25, 50, 75, 90 y 95 de la variable *salario* (salario actual) del archivo *Datos de empleados*:

```
EXAMINE VAR = salario  /PERCENTILES (5, 10, 25, 50, 75, 90, 95) HAVERAGE.
EXAMINE VAR = salario  /PERCENTILES (5, 10, 25, 50, 75, 90, 95) WAVERAGE.
EXAMINE VAR = salario  /PERCENTILES (5, 10, 25, 50, 75, 90, 95) ROUND.
EXAMINE VAR = salario  /PERCENTILES (5, 10, 25, 50, 75, 90, 95) EMPIRICAL.
EXAMINE VAR = salario  /PERCENTILES (5, 10, 25, 50, 75, 90, 95) AEMPIRICAL.
```

La Tabla 4.14 ofrece los resultados obtenidos con cada método de cálculo. Estos resultados permiten apreciar el grado de parecido entre los cinco métodos.

Tabla 4.14. Valores de algunos percentiles de la variable *salario actual*

Métodos de cálculo	*Percentiles*						
	5	10	25	50	75	90	95
Haverage	19200	21000	24000	28875	37162	59700	70218
Waverage	19200	21500	24000	28800	36825	59390	70000
Round	19200	21000	24000	28800	37050	59400	70000
Empirical	19200	21000	24000	28800	37050	59400	70000
Aempirical	19200	21000	24000	28875	37050	59400	70000

Ejercicios

4.1. El objetivo del análisis descriptivo es el de formarse una idea lo más exacta posible acerca de las características de un conjunto de datos. Esto se consigue prestando atención a tres propiedades de los datos: su centro, su dispersión y la forma de su distribución. En este ejercicio y en los dos siguientes vamos a intentar formarnos una idea lo más exacta posible acerca de las puntuaciones obtenidas al aplicar una prueba de selección (Y) a los 20 candidatos a un puesto de trabajo:

Candidatos	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Y	2	7	10	11	13	14	15	15	16	16	17	17	17	18	18	18	18	19	19	20

(las puntuaciones ya están ordenadas de menor a mayor para facilitar los cálculos). Se trata de una variable cuantitativa de la que hay que intentar formarse una idea lo más exacta posible. Para ello, vamos a comenzar averiguando cuál es su centro calculando:

a. La media aritmética.
b. La media recortada al 5%.
c. La mediana.
d. La trimedia.
e. Ofrecer una estimación del centro de la variable basada en los estadísticos calculados.

4.2. El centro de una variable no nos sirve de mucho si no va acompañado de alguna medida de dispersión. El grado de dispersión ayuda a precisar en qué medida el centro de la variable es un buen representante del resto de los valores. Con las puntuaciones de los 20 candidatos del ejercicio anterior, calcular:

a. La amplitud total.
b. La amplitud intercuartil.
c. La varianza y la desviación típica.
d. El coeficiente de variación basado en la media.
e. Comentar los resultados obtenidos.

4.3. Siguiendo con los datos del ejercicio 4.1, vamos a ocuparnos, por último, de la forma de la distribución. Es la mejor manera de detectar la presencia de asimetrías y de posibles anomalías en los datos. Obtener:

a. El histograma (sin agrupar datos en categorías).
b. El diagrama de caja.
c. El índice de asimetría.
d. El índice de curtosis.
e. Los errores típicos de los índices de asimetría y curtosis.
f. El cociente entre los índices de asimetría y curtosis y sus respectivos errores típicos.
g. Comentar los resultados obtenidos.

4.4. El siguiente conjunto de puntuaciones procede de una muestra de 10 pacientes a los que se ha administrado la escala de depresión de Hamilton antes (Y_{antes}) y después ($Y_{\text{después}}$) de recibir un tratamiento antidepresivo:

Pacientes	1	2	3	4	5	6	7	8	9	10
Y_{antes}	9	8	21	14	3	17	14	33	22	19
$Y_{\text{después}}$	10	2	11	6	3	10	3	15	8	12

Calcular, en ambos momentos:

a. La media aritmética.
b. La desviación típica.
c. Los tres cuartiles del momento *antes*.

4.5. Vamos a repasar algunas propiedades de la media aritmética utilizando los datos del ejercicio anterior (el lector poco familiarizado con las reglas básicas del sumatorio, puede revisar el Apéndice 4):

a. Si a las puntuaciones del momento *antes* se les resta su media, ¿cuánto vale la suma de las diferencias?
b. Si las puntuaciones del momento antes se multiplican por 10 y a los productos resultantes se les suma 5, ¿cuánto vale la media de las nuevas puntuaciones?
c. Si calculamos las diferencias entre los momentos *antes* y *después*, ¿cuánto vale la media de esas diferencias?
d. Si las puntuaciones del momento *después* se multiplican por una constante, le media de las nuevas puntuaciones vale 24. ¿Cuál es esa constante?
e. Al sumar una constante a las puntuaciones del momento *después*, la media de las nuevas puntuaciones vale 13. ¿Qué constante se ha sumado?

4.6. La siguiente tabla ofrece las calificaciones obtenidas por 30 estudiantes en una asignatura cualquiera (las calificaciones ya están ordenadas de menor a mayor para facilitar los cálculos):

Sujetos	1º	2º	3º	4º	5º	6º	7º	8º	9º	10º	11º	12º	13º	14º	15º
Calificaciones	2,5	2,8	3,2	3,5	3,8	4	4,5	4,7	5	5	5	5	5,2	5,5	5,5
Sujetos	16º	17º	18º	19º	20º	21º	22º	23º	24º	25º	26º	27º	28º	29º	30º
Calificaciones	5,5	5,8	6	6,2	6,4	7	7,5	7,6	7,8	8	8	8,2	8,6	9	9,5

a. Si el profesor de la asignatura decide que haya un 25% de suspensos, un 40% de aprobados, un 25% de notables y un 10% de sobresalientes, ¿dónde debe poner los puntos de corte?

b. Entre qué calificaciones se encuentra el 50% de los casos centrales.

4.7. Un educador ha administrado en tres aulas de 4º de enseñanza secundaria una prueba de inteligencia general para obtener el cociente intelectual (CI) de los estudiantes. En la siguiente tabla se ofrece, para cada una de las tres aulas, el cociente intelectual medio obtenido y el número de estudiantes:

Aulas	A	B	C
Nº de estudiantes	22	30	28
CI medio	98	103	100

a. ¿Cuánto vale el CI medio de todos los estudiantes tomados juntos?

b. Al añadir una cuarta aula con un CI medio de 105, el CI medio de todos los estudiantes ha subido a 102. ¿Cuántos estudiantes tiene la cuarta aula?

4.8. Una muestra aleatoria de sujetos obtiene una media de 60 en una variable cuyo rango de posibles valores va de 0 a 100. Si comprobamos que por encima de la puntuación 60 se encuentra el 60% de los sujetos:

a. ¿Es posible saber si la distribución es simétrica o asimétrica?

b. Si la distribución es asimétrica, ¿es asimétrica positiva o asimétrica negativa?

4.9. La media aritmética de dos números vale 10 y uno de ellos es tres veces mayor que el otro. ¿Cuáles son esos dos números?

4.10. Las puntuaciones de una variable se han multiplicado por la constante 10 y a los productos resultantes se les ha sumado la constante 5. Sabemos que la desviación típica de las puntuaciones transformadas vale 150. ¿Cuánto vale la desviación típica de las puntuaciones originales?

4.11. Un grupo de personas obtiene una varianza de 10 en la variable X y una varianza de 20 en la variable Y. ¿Puede afirmarse que el grado de dispersión de la variable Y es mayor que el de la variable X?

4.12. El percentil 25 de un conjunto de puntuaciones vale 12, la mediana vale 15 y el percentil 75 vale 20. La distribución de las puntuaciones ¿es simétrica, asimétrica negativa o asimétrica positiva?

4.13. La varianza de las puntuaciones $Y = \{0, 1, 3, 5, Y_5, 12\}$ vale 22. Sabiendo que el coeficiente de variación centrado en la media vale 93,81, calcular el valor de la puntuación que falta (Y_5) y el percentil 25 de las 6 puntuaciones.

4.14. En un *pub* escocés reza la siguiente leyenda: "Cuando un escocés emigra a Inglaterra mejora el cociente intelectual de los ingleses... y el de los escoceses". ¿Es esto posible? Razona la respuesta.

4.15. Una propiedad de la media que no hemos mencionado es que la suma de las desviaciones de la media al cuadrado es menor que la suma de las desviaciones respecto de cualquier otro valor, es decir:

$$\sum_i (Y_i - \bar{Y}) < \sum_i (Y_i - c) \qquad (\text{para } c \neq \bar{Y})$$

Imaginemos que una variable medida en 20 sujetos tiene una media de 10. Si la suma de las desviaciones de la media al cuadrado vale 130, ¿cuánto vale la suma de las desviaciones de la puntuación 15?

Soluciones

4.1. *a*. $\bar{Y} = (2+7+10+11+\cdots+18+19+19+20)\,/\,20 = 300\,/\,20 = 15$.

b. El 5% de 20 casos es 0,05 (20) = 1. Por tanto, la media recortada al 5% es la media aritmética tras eliminar dos casos (un caso de la parte inferior de la distribución y otro caso de la parte superior). Es decir,

$\bar{Y}_{5\%} = (7+10+11+\cdots+18+19+19)\,/\,18 = 278\,/\,18 = 15{,}44$.

c. Puesto que $i = (n+1)/2 = (20+1)\,/\,2 = 10{,}5$ la mediana hay que calcularla tomando como referencia las posiciones 10ª y 11ª, es decir, las puntuaciones 16 y 17:

$Mdn_Y = (1-0{,}5)16 + (0{,}5)17 = 16{,}5$.

d. Para poder calcular la trimedia necesitamos los tres cuartiles. Pero el segundo cuartil ya lo hemos calculado en el apartado anterior: $Q_2 = Mdn_Y = 16{,}5$.

- $Q_1 = P_{25}$: $i = k(n+1)/100 = 25\,(20+1)/100 = 5{,}25$. Hay que trabajar con las posiciones 5ª y 6ª, es decir, con las puntuaciones 13 y 14: $P_{25} = (1-0{,}25)13 + (0{,}25)14 = 13{,}25$.
- $Q_3 = P_{75}$: $i = k(n+1)/100 = 75\,(20+1)/100 = 15{,}75$. Hay que trabajar con las posiciones 15ª y 16ª, es decir, con las puntuaciones 18 y 18. Y como ambas puntuaciones son iguales: $P_{25} = 18$.

Por tanto, $\bar{Q} = (Q_1 + 2Q_2 + Q_3)/4 = (13{,}25 + 2(16{,}5) + 18)\,/\,4 = 16{,}06$.

e. Los estadísticos calculados ofrecen valores comprendidos entre 15 y 16,5. El más resistente (la mediana) vale 16,5. El menos resistente (la media) vale 15. Por tanto, el centro de la variable se encuentra en torno a 16 puntos. El hecho de que la media sea menor que la mediana permite anticipar que la distribución es asimétrica negativa. Enseguida veremos cuál es el grado de asimetría.

4.2. *a*. $A_T = 20 - 2 = 18$.

b. $A_{IQ} = 18 - 13{,}25 = 4{,}75$ (los cuartiles 1º y 3º se han calculado en el ejercicio anterior).

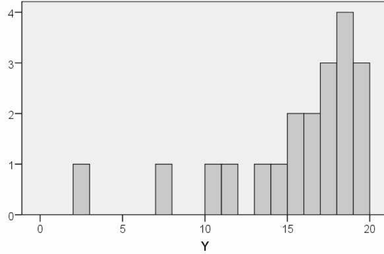
c. $S_Y^2 = [(2-15)^2 + (7-15)^2 + (10-15)^2 + \cdots + (19-15)^2 + (19-15)^2 + (20-15)^2]\,/\,19 = 20{,}316$.

$S_Y = (20{,}316)^{1/2} = 4{,}51$.

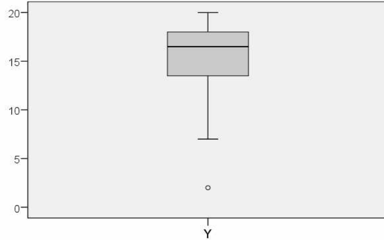
d. $CV_{\text{media}} = 100\,(4{,}51/15) = 30{,}07$.

e. La amplitud total no supera 4 veces a la intercuartil; esto significa que no existe mucha dispersión (éste es el criterio en el que se basan, por ejemplo, los bigotes de los diagramas de caja). Además, el coeficiente de variación es aproximadamente un tercio de la media, lo cual indica, de nuevo, que se trata de un conjunto de puntuaciones no muy dispersas.

4.3. *a*. Histograma (sin agrupar datos en categorías):



b. Diagrama de caja:



c. $M_3 = (2-15)^3+(7-15)^3+(10-15)^3+\cdots+(19-15)^3+(19-15)^3+(20-15)^3 = -2.520.$
$g_1 = [nM_3]/[(n-1)(n-2)S_Y^3] = 20(-2.520)/[19(18)(4{,}51)^3] = -1{,}61.$

d. $M_2 = (2-15)^2+(7-15)^2+(10-15)^2+\cdots+(19-15)^2+(19-15)^2+(20-15)^2 = 386.$
$M_4 = (2-15)^4+(7-15)^4+(10-15)^4+\cdots+(19-15)^4+(19-15)^4+(20-15)^4 = 35.066.$
$g_2 = [n(n+1)M_4 - 3(n-1)M_2^2]/[(n-1)(n-2)(n-3)S_Y^4] =$
$= [20(21)(35.066) - 3(19)(386)^2]/[19(18)(17)(4{,}51)^4] = 2{,}59.$

e. $S_{g_1} = \sqrt{[6n(n-1)]/[(n+1)(n-2)(n+3)]} = \sqrt{[6(20)(19)/[21(18)(17)]} = 0{,}51.$
$S_{g_2} = \sqrt{[4(n^2-1)S_{g_1}^2]/[(n-3)(n+5)]} = \sqrt{[4(20^2-1)(0{,}51^2)]/[17(25)]} = 0{,}99.$

f. $Z_{g_1} = g_1/S_{g_1} = -1{,}61/0{,}51 = -3{,}16.$
$Z_{g_2} = g_2/S_{g_2} = 2{,}59/0{,}99 = 2{,}62.$

g. Tanto el histograma como el diagrama de caja muestran una distribución asimétrica negativa. Y el cociente entre el índice de asimetría y su error típico (que es menor que −2) confirma la presencia de asimetría. Los casos que se alejan del centro por la izquierda hacen que en esa zona haya más casos (leptocurtosis) de los que habría en una distribución mesocúrtica.

Resumiendo: las 20 puntuaciones obtenidas en la prueba de selección tienen las siguientes características: un centro situado en torno a 16, sin excesiva dispersión y con evidente asimetría negativa.

4.4. *a*. $\bar{Y}_{\text{antes}} = (9+8+21+14+3+17+14+33+22+19)/10 = 160/10 = 16.$
$\bar{Y}_{\text{después}} = (10+2+11+6+3+10+3+15+8+12)/10 = 80/10 = 8.$

b. $S_{Y_{\text{antes}}}^2 = [(9-16)^2+(8-16)^2+\cdots+(22-16)^2+(19-16)^2]/9 = 650/9 = 72{,}222.$
$S_{Y_{\text{antes}}} = (72{,}222)^{1/2} = 8{,}50.$

$S^2_{Y_{\text{después}}} = [(10-8)^2+(2-8)^2+(11-8)^2+\cdots+(8-8)^2+(12-8)^2]/9 = 172/9 = 19{,}111.$

$S_{Y_{\text{después}}} = (19{,}111)^{1/2} = 4{,}37.$

c. Y_{antes}: 3, 8, 9, 14, 14, 17, 19, 21, 22, 33 (puntuaciones ordenadas).

- $Q_1 = P_{25}$: $i = k(n+1)/100 = 25(10+1)/100 = 2{,}75$. Hay que trabajar con las posiciones 2ª y 3ª, es decir, con las puntuaciones 8 y 9: $P_{25} = (1-0{,}75)8 + (0{,}75)9 = 8{,}75$.
- $Q_2 = P_{50}$: $i = k(n+1)/100 = 50(10+1)/100 = 5{,}50$. Hay que trabajar con las posiciones 5ª y 6ª, es decir, con las puntuaciones 14 y 17: $P_{50} = (1-0{,}50)14 + (0{,}50)17 = 15{,}50$.
- $Q_3 = P_{75}$: $i = k(n+1)/100 = 75(10+1)/100 = 8{,}25$. Hay que trabajar con las posiciones 8ª y 9ª, es decir, con las puntuaciones 21 y 22: $P_{75} = (1-0{,}25)21 + (0{,}25)22 = 21{,}25$.

4.5. *a*. $\sum_i (Y_i - \bar{Y}) = \sum_i Y_i - \sum_i \bar{Y} = n\bar{Y} - n\bar{Y} = 0.$

$\sum_i (Y_i - \bar{Y}) = (9-16) + (8-16) + (21-16) + \cdots + (22-16) + (19-16) = 0.$

b. $X = a + bY; \quad \bar{X} = \frac{1}{n}\sum_i (a + bY_i) = \frac{1}{n}na + \frac{1}{n}b\sum_i Y_i = a + b\bar{Y}.$

$\bar{X} = a + b\bar{Y} = 5 + 10(16) = 165.$

c. $X = Y_{\text{antes}} - Y_{\text{después}}; \quad \bar{X} = \frac{1}{n}\sum_i (Y_{\text{antes}} - Y_{\text{después}}) = \frac{1}{n}\sum_i Y_{\text{antes}} - \frac{1}{n}\sum_i Y_{\text{después}} = \bar{Y}_{\text{antes}} - \bar{Y}_{\text{después}}.$

$\bar{X} = \bar{Y}_{\text{antes}} - \bar{Y}_{\text{después}} = 16 - 8 = 8.$

d. $X = bY_{\text{después}}, \quad \bar{X} = 24.$

$\bar{X} = b\bar{Y}_{\text{después}} = b(8) = 24 \rightarrow b = 24/8 = 3.$

e. $X = a + Y_{\text{después}}, \quad \bar{X} = 24.$

$\bar{X} = a + \bar{Y}_{\text{después}} = a + 8 = 13 \rightarrow a = 13 - 8 = 5.$

4.6. *a*. Los puntos de corte deben colocarse en las calificaciones que correspondan al percentil 25, al percentil $25 + 40 = 65$ y al percentil $65 + 25 = 90$:

P_{25}: $i = k(n+1)/100 = 25(30+1)/100 = 7{,}75$. Hay que trabajar con las posiciones 7ª y 8ª, es decir, con las puntuaciones 4,5 y 4,7: $P_{25} = (1-0{,}75)4{,}5 + (0{,}75)4{,}7 = 4{,}65$.

P_{65}: $i = k(n+1)/100 = 65(30+1)/100 = 20{,}15$. Hay que trabajar con las posiciones 20ª y 21ª, es decir, con las puntuaciones 6,4 y 7,0: $P_{65} = (1-0{,}15)6{,}4 + (0{,}15)7{,}0 = 6{,}49$.

P_{90}: $i = k(n+1)/100 = 90(30+1)/100 = 27{,}90$. Hay que trabajar con las posiciones 27ª y 28ª, es decir, con las puntuaciones 8,2 y 8,6: $P_{90} = (1-0{,}90)8{,}2 + (0{,}90)8{,}6 = 8{,}56$.

b. Necesitamos conocer los percentiles 25 y 75. El percentil 25 vale 4,65 (lo hemos calculado en el apartado anterior). P_{75}: $i = k(n+1)/100 = 75(30+1)/100 = 23{,}25$. Hay que trabajar con las posiciones 23ª y 24ª, es decir, con las puntuaciones 7,6 y 7,8: $P_{75} = (1-0{,}25)7{,}6 + (0{,}25)7{,}8 = 7{,}65$. Por tanto, el 50% de los casos centrales se encuentra entre las calificaciones 4,65 y 7,65.

4.7. *a*. Media ponderada:

$$\bar{Y}_{\text{total}} = \frac{n_A\bar{Y}_A + n_B\bar{Y}_B + n_C\bar{Y}_C}{n_A + n_B + n_C} = \frac{22(98) + 30(103) + 28(100)}{22 + 30 + 28} = \frac{8.046}{80} = 100{,}575.$$

b. $$102 = \frac{8.046 + 105n_D}{80 + n_D} \rightarrow n_D = \frac{102(80) - 8.046}{105 - 102} = 38.$$

4.8. *a*. Más de la mitad de los casos (el 60%) obtiene puntuaciones mayores que la media. Por tanto, la media difiere de la mediana. Y esto implica que la distribución es asimétrica.

b. Por encima de la media quedan más casos (60%) que por encima de la mediana (50%). Por tanto la media es menor que la mediana. Esto es lo que ocurre en una distribución asimétrica negativa, donde los valores que se alejan por la parte baja de la distribución hacen que la media se desplace hacia esa zona.

4.9. $Y_1 + Y_2 = 2(10) = 20 \quad \rightarrow \quad Y_1 + 3Y_1 = 20 \quad \rightarrow \quad Y_1 = 20/4 = 5.$

$Y_2 = 3Y_1 = 3(5) = 15.$

4.10. La desviación típica no se altera si a las puntuaciones originales se les suma una constante, pero queda multiplicada por la constante multiplicada. Por tanto, la desviación típica de las puntuaciones originales vale 150/10 = 15.

4.11. No. El valor de la varianza no sólo depende del grado de dispersión de una variable, sino de su métrica. Por tanto, si no se conoce la métrica de las variables no es posible afirmar cuál de ellas es más dispersa. Si la altura se mide en metros y las puntuaciones obtenidas tienen una varianza de 0,20, esas mismas puntuaciones expresadas en centímetros tendrán una varianza de $100^2(0{,}20) = 2.000$.

4.12. La distribución es asimétrica porque los percentiles 25 y 75 no se encuentran a la misma distancia de la mediana. Y la asimetría es positiva porque el percentil 75 se encuentra más alejado de la mediana (5 puntos) que el percentil 25 (3 puntos).

4.13. $CV_{\text{media}} = 100\,S_Y/\bar{Y} \quad \rightarrow \quad 100\sqrt{22}/\bar{Y} = \sqrt{22}/\bar{Y} \quad \rightarrow \quad \bar{Y} = 100\sqrt{22}/93{,}81 = 5.$

$\sum_i \bar{Y} = n\bar{Y} = 6(5) = 30 \quad \rightarrow \quad Y_5 = 30 - 0 - 1 - 3 - 5 - 12 = 9.$

P_{25}: $i = k(n+1)/100 = 25(6+1)/100 = 1{,}75$. Hay que trabajar con las posiciones 1ª y 2ª, es decir, con las puntuaciones 0 y 1: $P_{25} = (1 - 0{,}75)0 + (0{,}75)1 = 0{,}75$.

4.14. Esto sólo ocurrirá si el CI del escocés que emigra a Inglaterra es menor que el CI medio escocés y mayor que el CI medio inglés. Y esto sólo será posible si el CI medio escocés es mayor que el CI medio inglés. Por tanto, sí es posible. Pero obviamente se trata de una broma escocesa en la que se da por hecho que el CI medio de los ingleses es menor que el de los escoceses y que si un escocés decide emigrar a Inglaterra es porque su CI está por debajo del CI medio escocés.

4.15. $\sum_i (Y_i - c)^2 \quad (\text{con } c \neq \bar{Y}).$

Hagamos: $c = \bar{Y} + a$.

$$\sum_i (Y_i - c)^2 = \sum_i [Y_i - (\bar{Y} + a)]^2 = \sum_i [(Y_i - \bar{Y}) + a]^2 =$$
$$= \sum_i (Y_i - \bar{Y})^2 + na^2 + 2a\sum_i (Y_i - \bar{Y}) = \sum_i (Y_i - \bar{Y})^2 + na^2.$$

Es decir, la suma de las nuevas desviaciones al cuadrado es igual a la suma de las desviaciones de la media al cuadrado más el tamaño de la muestra multiplicado por el cuadrado de *a* (y *a* no es otra cosa que la diferencia entre la media y *c*). En el ejemplo:

$$\sum_i (Y_i - c)^2 \sum_i (Y_i - \bar{Y})^2 + na^2 \quad \rightarrow \quad \sum_i (Y_i - c)^2 = 130 + 20(5) = 230.$$

5

Puntuaciones típicas y curva normal

En el recorrido hecho en los dos capítulos anteriores por las herramientas descriptivas hemos puesto el énfasis en tres aspectos básicos: el centro, la dispersión y la forma de la distribución. Pero también hemos señalado que hay otro tipo de información que puede resultar interesante. Con variables dicotómicas, por ejemplo, hemos visto que la distribución de probabilidad binomial permite obtener información adicional muy valiosa. Pues bien, con variables cuantitativas ocurre algo parecido; en concreto, hay dos herramientas estadísticas que ayudan de forma importante a completar la información que ofrecen el centro, la dispersión y la forma de la distribución. Nos estamos refiriendo a un tipo particular de transformación de los datos que suele recibir el nombre de ***tipificación*** y a una distribución teórica de probabilidad que sirve como referente del comportamiento de muchas de las variables cuantitativas que suele interesar estudiar: la ***curva normal***.

Puntuaciones típicas (Z)

Entre las tareas más habituales que tiene que abordar un analista de datos se encuentra la de realizar comparaciones. Estas comparaciones se suelen hacer entre grupos de puntuaciones tomando algún promedio como referente para la comparación. Ahora bien, aunque estas comparaciones entre promedios constituyen uno de los principales objetivos del análisis de datos en la fase inferencial, en la fase descriptiva puede interesar realizar comparaciones, no entre promedios, sino entre puntuaciones individuales.

Consideremos un grupo de sujetos en el que se ha medido la altura y el peso. ¿Es posible comparar una altura de 175 cm con un peso de 80 kg? En principio, no, pues se trata de variables con ***diferente métrica*** (diferentes unidades de medida: *cm* y *kg*). De hecho, si la altura se midiera en metros el resultado de la comparación sería completamente distinto.

Pero ¿qué ocurre si las variables tienen la misma métrica? La Tabla 5.1 muestra las calificaciones de 12 sujetos en tres asignaturas: lengua, matemáticas y filosofía. Las calificaciones de lengua y matemáticas tienen *distinto centro pero la misma dispersión*. ¿Qué puede decirse

de una calificación de 4? La respuesta depende de la asignatura: en lengua se trata de la peor calificación; en matemáticas, de una calificación intermedia (coincide exactamente con la media del grupo). Por tanto, el significado de una misma calificación depende de la asignatura. Esto se debe a que las distribuciones de las dos asignaturas tienen ***diferente centro***.

Las calificaciones de lengua y filosofía tienen *el mismo centro pero diferente dispersión*. ¿Qué puede decirse de una calificación de 8? De nuevo la respuesta depende de la asignatura que se tome como referente: en lengua se trata de la calificación más alta; en filosofía, de una calificación que tiene varias calificaciones por encima. También ahora el significado de una misma calificación depende de la asignatura. Pero, esta vez, se debe a que las distribuciones de las dos asignaturas tienen ***diferente dispersión***.

Tabla 5.1. Calificaciones de un grupo de 12 estudiantes en 3 asignaturas

Asignaturas	*Calificaciones (puntuaciones directas)*												$\bar{Y}_j$	S_{Y_j}
Lengua	4,0	4,8	5,0	5,2	5,6	6,0	6,0	6,3	6,7	7,0	7,4	8,0	6,0	1,16
Matemáticas	2,1	2,8	3,0	3,2	3,4	3,8	4,0	4,6	4,9	5,0	5,2	6,0	4,0	1,16
Filosofía	1,8	3,5	4,5	4,5	5,0	4,8	6,0	8,0	8,2	8,2	8,5	9,0	6,0	2,33

Estos sencillos ejemplos sirven para llamar la atención sobre una cuestión importante: las puntuaciones de distribuciones distintas no admiten una comparación directa a no ser que esas distribuciones tengan el *mismo centro*, la *misma dispersión* y la *misma métrica*. Sin embargo, esto no significa que haya que renunciar a comparar las puntuaciones que no cumplan esos tres requisitos. Es más, para abordar esta tarea disponemos de dos estrategias distintas.

La primera de ellas ya ha sido presentada en el capítulo anterior; nos referimos a los *percentiles*. Recordemos que los percentiles permiten conocer la posición relativa que ocupa cada puntuación dentro de su distribución tomando como base de cálculo el *porcentaje de casos acumulados*. Con esta estrategia, las puntuaciones originales o directas (cualquiera que sea su métrica) se transforman en posiciones relativas que pasan a tener el mismo centro (50), la misma dispersión (0 – 100) y la misma métrica (unidades porcentuales). Por tanto, con los percentiles es posible realizar comparaciones entre puntuaciones de distintos grupos o variables porque la comparación se basa, no en las puntuaciones directas, sino en las posiciones relativas que esas puntuaciones ocupan en sus respectivos grupos o variables.

El problema de los percentiles es que únicamente aprovechan información ordinal, lo cual implica que son insensibles a cambios importantes en los datos. Por ejemplo, si la calificación 4 en matemáticas se cambia por un 10 (un cambio importante que altera tanto el centro como la dispersión del conjunto de calificaciones), los percentiles correspondientes a las primeras cinco calificaciones de esa asignatura no se alteran.

Esta limitación inherente a los percentiles puede resolverse utilizando una estrategia basada, no en el porcentaje de casos acumulados, sino en las *distancias al centro*, en concreto, en las *distancias a la media* (es decir, en las *puntuaciones diferenciales* o de *desviación*). Ahora bien, las distancias a la media, en bruto, no parece que ayuden a resolver el problema. Ciertamente permiten comparar puntuaciones directas con *distinto centro* pues, en las distancias, las medias quedan igualadas a cero. Pero no resuelven el problema del *diferente grado de dispersión* (pues la calificación 8, por ejemplo, se encuentra a 2 puntos de la media tanto en lengua como en filosofía a pesar de que en lengua ocupa la posición más alta y en filosofía ocupa una posición más bien intermedia) ni el problema de la *diferente métrica* (pues las distancias

a la media están expresadas en la misma métrica que las puntuaciones directas: la distancia de una altura a su centro sigue expresada en *cm* y la de un peso a su centro en *kg*).

La solución pasa por transformar las distancias a la media en puntuaciones que, además del mismo centro (pues la media de las distancias a la media vale cero), también tengan la misma dispersión y la misma métrica. Pues bien, esto es justamente lo que hacen las **puntuaciones típicas** o **puntuaciones *Z*** simplemente dividiendo las distancias a la media entre la desviación típica:

$$Z = \frac{Y - \bar{Y}}{S_Y} \qquad [5.1]$$

Al dividir entre la desviación típica, las distancias a la media quedan expresadas en unidades de desviación típica (en *unidades de dispersión*) y con ello se obtienen unas nuevas puntuaciones que cumplen los tres requisitos necesarios para comparar puntuaciones de distintos grupos o variables, es decir, se obtienen puntuaciones con el mismo centro, la misma dispersión y la misma métrica. En efecto, con la tipificación se obtienen:

1. *Puntuaciones con el mismo centro*: cualquiera que sea la media de la variable original, la media de las nuevas puntuaciones *Z* vale cero, pues no son más que transformaciones lineales de puntuaciones cuya media ya vale cero[1] (recuérdese que la suma de las desviaciones de la media vale cero; ver ecuación [3.3]). Esto significa que las puntuaciones *Z* positivas corresponden a puntuaciones directas mayores que la media y las negativas a puntuaciones directas menores que la media; a la media de las puntuaciones directas le corresponde una puntuación *Z* de cero.
2. *Puntuaciones con la misma dispersión*: cualquiera que sea el grado de dispersión de la variable original, la varianza y la desviación típica[2] de las nuevas puntuaciones *Z* vale uno.
3. *Puntuaciones con la misma métrica*: cualquiera que sea la métrica original de la variable tipificada, la métrica de las nuevas puntuaciones *Z* cambia a unidades de desviación típica. Un cambio de una unidad en la nueva métrica siempre está indicando un cambio de una desviación típica[3].

Por tanto, las puntuaciones *Z* siempre tienen una media de 0 y una desviación típica de 1 (sin importar la métrica de las puntuaciones originales o directas). Y, aunque, en teoría, no tienen límite mínimo y máximo, sabemos que suelen tomar valores comprendidos ente −3 y 3. Independientemente de las características de la distribución original, las puntuaciones *Z* mayores que 3 o menores que −3 suelen estar delatando casos atípicos (casos cuyas puntuaciones se encuentran excesivamente alejadas del centro).

---

[1] $\bar{Z} = \sum (Y - \bar{Y})/(nS_Y) = 0/(nS_Y) = 0$ [5.2]

[2] $S_Z^2 = \frac{\sum (Z - \bar{Z})^2}{n-1} = \frac{\sum Z^2}{n-1} = \frac{\sum (Y - \bar{Y})^2}{S_Y^2\,(n-1)} = \frac{S_Y^2}{S_Y^2} = 1$ [5.3]

[3] No debe confundirse una *puntuación típica* (*Z*) con una *desviación típica* (S_Y). Una puntuación típica igual a 1 indica que la correspondiente puntuación directa se aleja una desviación típica de su media. Pero una puntuación típica es un valor *individual* (a cada puntuación directa le corresponde una puntuación típica), mientras que una desviación típica es un valor *grupal* (a todo el conjunto de puntuaciones le corresponde una única desviación típica).

La Tabla 5.2 muestra las puntuaciones típicas correspondientes a las calificaciones de la Tabla 5.1. Estas puntuaciones se han obtenido aplicando la ecuación [5.1]. Así, por ejemplo, a una calificación de 4 en lengua (asignatura con media = 6 y desviación típica = 1,16) le corresponde una puntuación típica de

$$Z_{(\text{lengua} = 4)} = \frac{4 - 6}{1{,}16} = -1{,}73$$

Tabla 5.2. Puntuaciones típicas (*Z*) correspondientes a las calificaciones de la Tabla 5.1

Asignaturas	*Calificaciones* (*puntuaciones típicas*)											
Lengua	−1,73	−1,04	−0,86	−0,69	−0,35	0	0	0,26	0,6	0,86	1,21	1,73
Matemáticas	−1,64	−1,04	−0,87	−0,69	−0,52	−0,17	0	0,52	0,78	0,87	1,04	1,73
Filosofía	−1,80	−1,07	−0,64	−0,64	−0,43	−0,51	0	0,86	0,94	0,94	1,07	1,29

Retomemos ahora las preguntas formuladas al principio de este apartado. ¿Qué podemos decir de una calificación de 4 en lengua y en matemáticas? Sabemos que la calificación 4 está en la parte inferior de las calificaciones de lengua y en la parte intermedia de las de matemáticas. Pues bien, las puntuaciones típicas reflejan correctamente estas posiciones relativas asignándole un valor de −1,73 en lengua y de 0 en matemáticas. Esto significa que la calificación 4 en lengua se encuentra a 1,73 desviaciones típicas por debajo de su media (el signo negativo indica que se trata de una puntuación menor que su media) y que la calificación 4 en matemáticas se encuentra justamente en el centro de su distribución (pues sólo a la media le corresponde una puntuación típica de cero).

Y ¿qué puede decirse de una calificación de 8 en lengua y en filosofía? La calificación 8 ocupa la posición más alta en lengua y una posición media-alta en filosofía. Las correspondientes puntuaciones típicas (1,73 en lengua y 0,86 en filosofía) reflejan correctamente estas posiciones relativas: aunque, en términos absolutos, la distancia de la calificación 8 a la media vale 2 puntos tanto en lengua como en filosofía, esa distancia es, en términos relativos, el doble de grande en lengua (1,73) que en filosofía (0,86).

De todo lo anterior cabe concluir que las puntuaciones típicas sirven para comparar puntuaciones individuales de distintos grupos y variables con bastante solvencia. Lo cual, no sólo permite constatar, como hemos visto, que dos puntuaciones directas iguales pueden ocupar posiciones relativas distintas, sino también que dos puntuaciones directas distintas pueden ocupar posiciones relativas iguales. Así, por ejemplo, a la calificación 5,2 en lengua le corresponde la misma puntuación típica (−0,69) que a la calificación 3,2 en matemáticas. Cuando se da esta circunstancia (puntuaciones directas con la misma puntuación *Z*), decimos que esas puntuaciones son **equivalentes**.

Conviene señalar que las puntuaciones *Z* no son la única forma de tipificación disponible. En estadística es habitual efectuar transformaciones para poder interpretar correctamente los datos, es decir, es habitual relativizar las puntuaciones originales, referirlas a *algo*, para dotarlas de sentido. El tipo de transformación aplicada depende de las características de los datos y del objetivo perseguido. Cuando, más adelante, estudiemos estadística inferencial, veremos que todos los procedimientos utilizados para realizar comparaciones y estudiar relaciones no son más que una forma de tipificación orientada a extraer de los datos el tipo de información que más interesa en cada caso.

Puntuaciones típicas y percentiles

Las dos estrategias propuestas para comparar puntuaciones entre distintos grupos o variables no son necesariamente equivalentes. Si dos puntuaciones directas, una de cada grupo o variable, tienen el mismo percentil, las correspondientes puntuaciones típicas sólo serán iguales cuando la forma de las dos distribuciones sea idéntica (aunque el centro sea distinto). Y si la forma de las dos distribuciones no es idéntica, dos puntuaciones directas, una de cada distribución, con distinto percentil pueden tener puntuaciones típicas iguales.

En este sentido, es importante señalar que, aunque la transformación en puntuaciones Z cambia tanto el centro como la dispersión de la variable original, *la forma de la distribución permanece inalterada* (esto no ocurre con los percentiles; la transformación en percentiles convierte la métrica original en ordinal –posiciones relativas– y, con ello, se altera por completo la forma de la distribución). Por tanto, si la distribución de la variable original es simétrica (o asimétrica, o leptocúrtica, etc.), la distribución de las correspondientes puntuaciones Z seguirá siendo simétrica (o asimétrica, o leptocúrtica, etc.). Esta propiedad posee importantes implicaciones cuya utilidad se comprenderá en los siguientes apartados.

Escalas derivadas

Las puntuaciones Z poseen otra interesante utilidad. Según acabamos de ver, toda variable cuantitativa, cualquiera que sea su métrica, puede ser tipificada en una nueva variable (puntuaciones Z) con media igual a 0 y desviación típica igual a 1. Pues bien, todo conjunto de puntuaciones Z puede, a su vez, ser transformado en un nuevo conjunto de puntuaciones con media y desviación típica conocidas.

Sabemos (ver [4.4.*c*]) que si a una variable se le multiplica y suma una constante, la media queda alterada; en concreto, si $X = a + bY$, entonces $\overline{X} = a + b\overline{Y}$. También sabemos (ver, en el capítulo anterior, el apartado *Comparación entre estadísticos de dispersión*) que la desviación típica no se altera por la constante sumada pero sí por la multiplicada; en concreto, si $X = a + bY$, entonces $S_X = |b|\,S_Y$. En consecuencia, la transformación

$$T = a + bZ \qquad [5.4]$$

tendrá una media $\overline{T} = a$ (pues la media de las puntuaciones Z vale cero) y una desviación típica $S_T = |b|$ (pues la desviación típica de las puntuaciones Z vale uno).

Esto significa que toda variable cuantitativa, cualquiera que sea su métrica, puede ser trasformada en otra variable equivalente con media y desviación típica conocidas (las puntuaciones transformadas son *equivalentes* a las originales porque a cada valor transformado le corresponde la misma puntuación típica que a su correspondiente valor original).

Este tipo de transformaciones sirven para que un conjunto de puntuaciones (por ejemplo, las puntuaciones de un test) puedan expresarse en la métrica deseada. Así, si las puntuaciones de un test oscilan entre 9 y 45, multiplicando las correspondientes puntuaciones típicas por 20 y sumándoles 50 se obtendría una escala derivada equivalente a la original con una media de 50 y un rango de puntuaciones comprendido entre, aproximadamente, 0 y 100 (lo cual suele entenderse mejor por las personas que responden al test o leen el informe de resultados).

La transformación puede hacerse utilizando cualquier valor para a y b, pero hay algunos valores que suelen utilizarse más que otros. Una de las transformaciones más conocidas consiste en sumar $a = 100$ y multiplicar $b = 15$ para obtener la escala de *CI* (cociente intelectual).

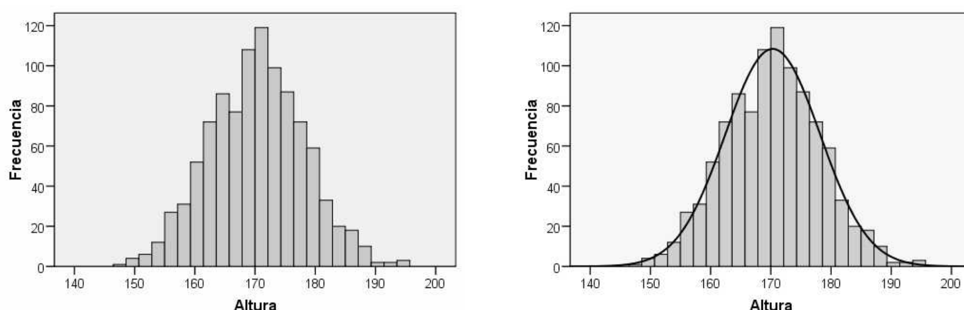
Curva normal

Ya hemos tenido ocasión de comprobar (ver Capítulo 3) que conocer la distribución teórica de probabilidad de una variable permite obtener información adicional a la que ofrecen su centro, dispersión y forma; en concreto, permite conocer las probabilidades asociadas a cada posible resultado muestral. Así, si se conocen las características poblacionales de una variable categórica (es decir, las frecuencias relativas asociadas a cada valor de la variable; por ejemplo, la proporción de fumadores), puede utilizarse la distribución binomial para conocer la probabilidad concreta asociada a cada resultado muestral (por ejemplo, la probabilidad de encontrar un determinado número de fumadores al extraer una muestra aleatoria de tamaño *n* de esa población).

Con variables cuantitativas ocurre algo parecido, pero el escenario es algo más complejo porque existen múltiples distribuciones posibles derivadas de combinar el grado de asimetría con el de curtosis. En un intento por abarcar todas esas posibilidades, se han diseñado muchas distribuciones teóricas que se utilizan como modelos para representar el comportamiento de los datos reales.

Ahora bien, aunque las distribuciones de las variables cuantitativas pueden adoptar formas muy diversas, muchas de las variables que medimos tienen una forma particular: la mayoría de los valores se encuentran próximos al centro de la distribución y van siendo menos frecuentes a medida que va aumentando la distancia al centro. Esto es lo que ocurre, por ejemplo, con la variable *altura* representada en los histogramas de la Figura 5.1. Este histograma se parece a una distribución teórica llamada **curva normal**. En una primera aproximación, la curva normal puede concebirse como una especie de histograma suavizado cuyas barras se han levantado sobre intervalos infinitamente pequeños (ver curva superpuesta en el segundo histograma de la Figura 5.1).

Figura 5.1. Histograma de la *altura* (izquierda) con curva normal superpuesta (derecha)



La curva normal es, probablemente, la distribución teórica más importante en estadística. Muchos de los procedimientos estadísticos que estudiaremos asumen que los datos proceden de poblaciones normales (tendremos ocasión de comprobar esto más adelante). Pero, además, la curva normal sirve como referente para describir cómo se distribuyen muchos de los datos que recogemos, pues muchos de los fenómenos naturales o sociales que interesa estudiar tienen distribuciones *empíricas* que se parecen a la distribución *teórica* normal. La justificación de esta regularidad se encuentra en el **teorema del límite central**, que, formulado en palabras, afirma lo siguiente:

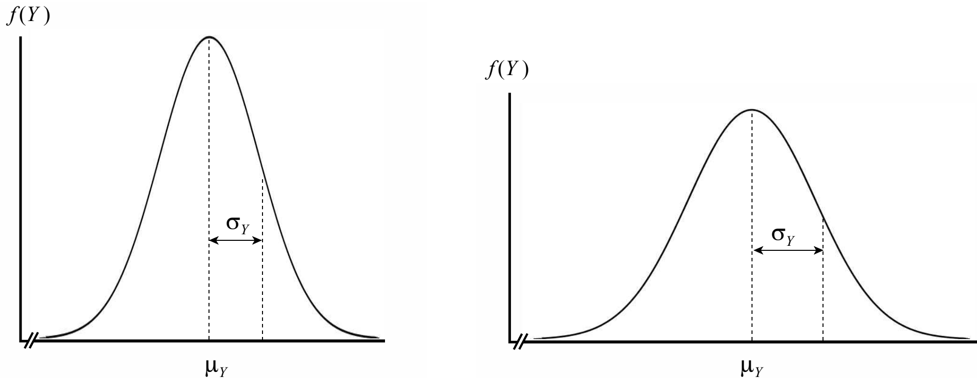
> *Si los datos que se recogen son debidos a la suma de cierto número de causas independientes entre sí, cada una con un efecto parcial, la distribución de los datos recogidos se asemejará tanto más a la curva normal cuantos más datos se recojan (cualquiera que sea la distribución original de esos efectos parciales y siempre que la desviación típica de estos efectos sea finita).*

La importancia de la distribución normal como referente teórico del comportamiento de muchos de las variables que interesa estudiar obliga a describir algunas de sus características. En la Figura 5.2 están representadas dos curvas normales. Aunque una es más estrecha y alta que la otra, ambas se generan con la misma ecuación:

$$f(Y) = \frac{1}{\sigma_Y \sqrt{2\pi}} \, e^{-(Y-\mu_Y)^2/(2\sigma_Y^2)} \qquad [5.5]$$

donde π y e son constantes conocidas ($\pi = 3{,}1416$; $e = 2{,}71828$), y μ_Y y σ_Y son, respectivamente, la media y la desviación típica de la distribución. La expresión $f(Y)$ se denomina *función de densidad* y permite generar la curva normal de Y asignando valores a μ_Y y σ_Y. Una vez obtenida la curva es posible conocer el área bajo la curva comprendida entre dos puntos (o entre un punto y el extremo izquierdo o el derecho) y, de esta forma, conocer la proporción de valores que se encuentran en cada porción de área.

Figura 5.2. Ejemplos de distribuciones (curvas) normales, con sus dos parámetros (μ_Y y σ_Y)



Para indicar que una variable se distribuye normalmente utilizaremos la siguiente expresión:

$$Y \sim N(\mu_Y, \sigma_Y) \qquad [5.6]$$

Cambiando los valores de μ_Y y σ_Y en [5.5] se obtienen distintas curvas normales. Por tanto, no existe una única curva normal, sino muchas. No obstante, todas ellas comparten las mismas características:

- Tienen un único máximo en μ_Y (por tanto, son unimodales).
- Tienen forma de campana (de ahí que también sean conocidas como *campanas de Gauss* –en referencia a su forma y a uno de sus autores). Esto implica que los valores centrales son más probables que los que se van alejando del centro.

- Son simétricas respecto al eje central situado en μ_Y. Por tanto, las diferentes medidas de tendencia central (media, mediana, moda, medias corregidas, etc.) coinciden.
- Son asintóticas respecto al eje de abscisas (por mucho que se extiendan, nunca llegan a tocarlo), por lo que los valores mínimo y máximo del eje de abscisas son $-\infty$ y $+\infty$. No obstante, sabemos que el 99,73% del área total se encuentra entre $\pm 3\sigma_Y$.
- Existen dos puntos de inflexión (aparte del que se produce en μ_Y) situados a una desviación típica (σ_Y) por encima y por debajo de μ_Y. En el punto $\mu_Y - \sigma_Y$ la curva pasa de ser cóncava hacia arriba a cóncava hacia abajo; en el punto $\mu_Y + \sigma_Y$ pasa de ser cóncava hacia abajo a cóncava hacia arriba.
- El área total bajo la curva vale 1. Todas las puntuaciones posibles se encuentran entre $-\infty$ y $+\infty$. Por tanto, la probabilidad de encontrar valores menores que $-\infty$ o mayores que $+\infty$ vale cero.
- Cualquier combinación lineal de variables (es decir, cualquier suma de variables cada una de ellas multiplicada por una constante) normalmente distribuidas también se distribuye según el modelo de probabilidad normal.

La mayor parte del trabajo con variables distribuidas normalmente consiste en hallar el área bajo la curva que queda por debajo o por encima de cada valor del eje de abscisas. Más concretamente, en hallar el tamaño relativo que cada porción de área representa respecto del área total (proporción de área). En este contexto, hablar de *proporción de área* es equivalente a hablar de *probabilidad*: en la primera curva normal de la Figura 5.2, la porción de área situada por debajo del valor μ_Y representa la proporción de valores menores que μ_Y; es decir, la probabilidad de encontrar valores por debajo de μ_Y. Estas probabilidades se obtienen a partir de las densidades que ofrece la función [5.5]. No obstante, para evitar este tipo de cálculos, se han construido tablas con las probabilidades (proporciones de área bajo la curva) ya calculadas (ver Tabla C del Apéndice final).

Ahora bien, estas tablas recogen las probabilidades de una curva normal muy especial llamada **curva normal tipificada** o *estandarizada*; es decir, una curva que tiene media 0 y desviación típica 1; lo cual se expresa mediante $N(0, 1)$. Por supuesto, el hecho de que la distribución normal tabulada tenga media 0 y desviación típica 1 no es un problema sino, de hecho, una ventaja, pues cualquier variable cuantitativa Y puede ser transformada en otra variable equivalente Z con media 0 y desviación típica 1 sin que se altere la forma de su distribución (a esta transformación la hemos llamado tipificación y, al resultado de la tipificación, puntuaciones típicas o puntuaciones Z; ver apartado anterior). Es decir,

$$\begin{array}{lll} \text{si} & Y \sim N(\mu_Y, \sigma_Y) & \\ \text{entonces} & Z = (Y - \mu_Y)/\sigma_Y \sim N(0, 1) & \end{array} \qquad [5.7]$$

Una vez que una puntuación directa ha sido transformada en puntuación típica, ya es posible conocer la probabilidad asociada a cada puntuación directa a partir de la probabilidad asociada a su correspondiente puntuación típica. La Figura 5.3 muestra la equivalencia existente entre las posiciones que ocupan las puntuaciones directas (gráficos de la parte superior) y las que ocupan sus correspondientes puntuaciones típicas (gráficos de la parte inferior). En estos gráficos se observa con claridad que la proporción de área que queda por debajo (o por encima) de una puntuación *directa* en una curva $N(\mu_Y, \sigma_Y)$ es exactamente la misma que queda por debajo (o por encima) de su correspondiente puntuación *típica* en la curva $N(0, 1)$. Y, co-

mo se está asumiendo que proporción es equivalente a probabilidad, lo que se está afirmando es que la probabilidad de encontrar valores menores (mayores) que una puntuación *directa* es exactamente la misma que la de encontrar valores menores (mayores) que su correspondiente puntuación *típica*. Es decir,

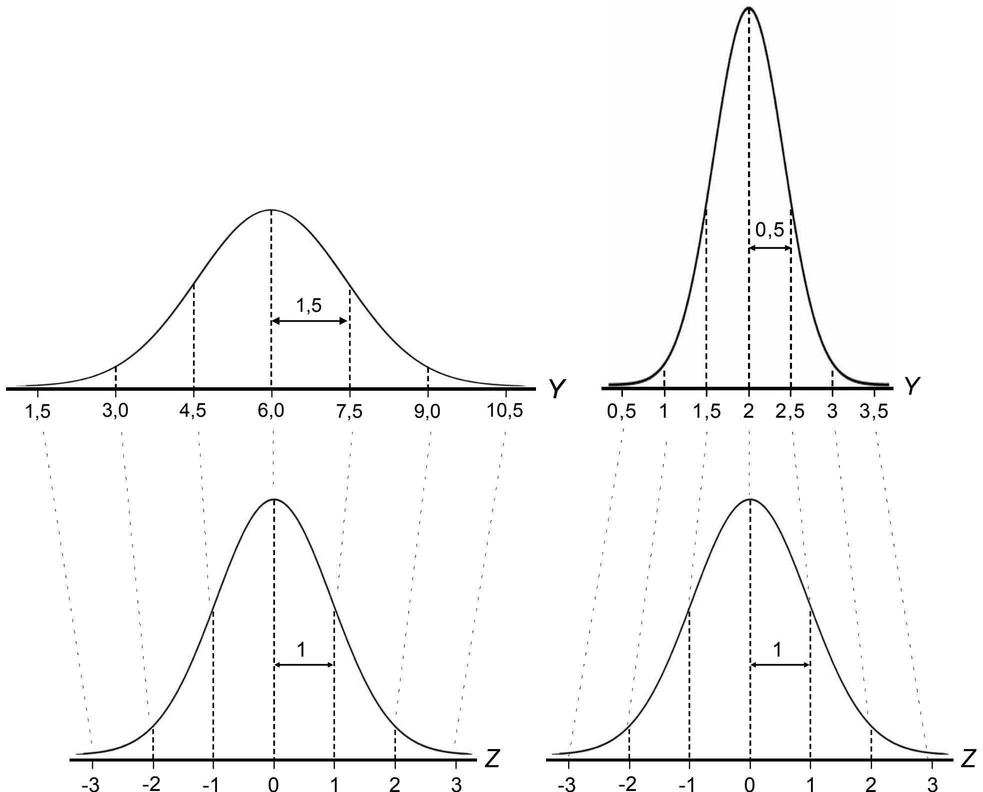
$$P(Y < Y_i) = P(Z < Z_i) \quad \text{y} \quad P(Y > Y_i) = P(Z > Z_i) \qquad [5.8]$$

Así las cosas, para conocer la probabilidad de encontrar valores menores (o mayores) que una puntuación directa es necesario: (1) transformar esa puntuación directa en puntuación típica y (2) utilizar la tabla de áreas bajo la curva normal (Tabla C del Apéndice final) para conocer la probabilidad buscada (en el siguiente apartado se explica cómo utilizar la Tabla C para encontrar estas probabilidades).

Para calcular, por ejemplo, la proporción de área que queda por debajo de la puntuación 4,5 en una distribución $N(6, 1{,}5)$, comenzamos transformando esa puntuación directa en puntuación Z siguiendo la ecuación [5.7]:

$$Z_{(Y=4,5)} = \frac{4{,}5 - 6}{1{,}5} = -1$$

Figura 5.3. Correspondencia entre curvas normales: los gráficos de la izquierda muestran la correspondencia entre la curva $N(6, 1{,}5)$ y la curva $N(0, 1)$; los de la derecha, entre la curva $N(2, 0{,}5)$ y la curva $N(0, 1)$



En la Figura 5.3 (izquierda) se observa que, efectivamente, a la puntuación $Y = 4{,}5$ le corresponde una puntuación $Z = -1$. Esto significa que la puntuación 4,5 en la distribución $N(6, 1{,}5)$ ocupa la misma posición relativa que la puntuación -1 en la distribución $N(0, 1)$. Y esto implica que la probabilidad de encontrar valores menores (o mayores) que 4,5 en su distribución es la misma que la de encontrar valores menores (o mayores) que -1 en la suya.

Una vez que sabemos que a la puntuación directa 4,5 le corresponde una puntuación típica de -1, podemos utilizar la *tabla de la curva normal* para averiguar qué proporción de área queda por debajo de una puntuación típica igual a -1.

Tabla de la curva normal

La tabla de la curva normal (Tabla C del Apéndice final) recoge la probabilidad (tamaño relativo del área bajo la curva) asociada a cada puntuación Z. Para leer correctamente la tabla hay que tener en cuenta, tal como indica el dibujo que aparece en la cabecera de la tabla, que los valores del interior de la tabla corresponden siempre a probabilidades *acumuladas*. En la primera columna de la tabla aparecen las puntuaciones Z (de $-3{,}0$ a $3{,}0$) con un solo decimal; las cabeceras de las columnas recogen el segundo decimal.

De acuerdo con esto, el primer valor de la primera fila de la tabla indica que la puntuación $Z = -3{,}00$ deja por debajo de sí (acumula) una probabilidad de 0,0013. El segundo valor de la tercera fila indica que la puntuación $Z = -2{,}81$ acumula una probabilidad de 0,0025.

La Figura 5.4 muestra tres ejemplos con las diferentes posibilidades que pueden surgir al trabajar con la tabla de la curva normal. El gráfico de la izquierda indica que la puntuación $Z = -0{,}82$ acumula una probabilidad de 0,2061. Puesto que se trata de una probabilidad acumulada, el valor se obtiene directamente de la tabla buscando la intersección entre la fila $-0{,}8$ y la columna 2. Este resultado se expresa de la siguiente manera[4]:

$$P(Z < -0{,}82) \;=\; F(-0{,}82) \;=\; 0{,}2061 \quad \rightarrow \quad Z_{0{,}2061} \;=\; -0{,}82$$

(F = probabilidad acumulada). Siempre que una puntuación Z aparezca con una probabilidad como subíndice, éste estará indicando la probabilidad que acumula esa puntuación Z; cuando el subíndice no se refiera a la probabilidad acumulada se colocará entre paréntesis.

El gráfico del centro indica que la puntuación $Z = 1{,}14$ deja por encima de sí una probabilidad de 0,1271. Este valor se obtiene restando de uno la probabilidad acumulada hasta 1,14. O, si se prefiere (más rápido), buscando directamente la probabilidad acumulada hasta $-1{,}14$; como la curva es simétrica respecto del eje que pasa por cero, la probabilidad acumulada hasta $-1{,}14$ es idéntica a la probabilidad que queda por encima de 1,14:

$$P(Z > 1{,}14) = P(Z < -1{,}14) \;=\; 0{,}1271 \quad \rightarrow \quad Z_{0{,}1271} \;=\; -1{,}14 \quad \rightarrow \quad Z_{0{,}8729} \;=\; 1{,}14$$

El gráfico de la derecha, por último, indica que la probabilidad comprendida entre $Z = 0{,}25$ y $Z = 1{,}14$ vale 0,2742. Esta probabilidad puede obtenerse de diferentes maneras, pero la más

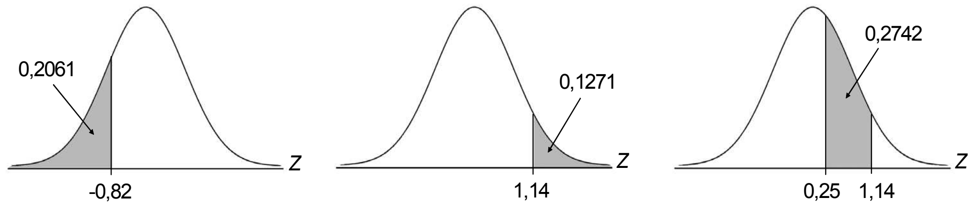
[4] En una distribución *continua* como la normal, la probabilidad asociada a un valor concreto es nula: $P(Y = Y_i) = 0$. Esto puede comprenderse fácilmente si se tiene en cuenta que toda el área bajo la curva (que se asume que vale 1) hay que repartirla entre los infinitos puntos que conforman el eje de abscisas. En el reparto, a cada uno de esos infinitos puntos le corresponde una cantidad de área infinitamente pequeña (nula). Esto no ocurre con distribuciones *discretas* como la binomial, donde cada valor tiene asociada una probabilidad concreta. Una consecuencia de esto es que, en una distribución continua, los signos "<" y ">" son equivalentes a los signos "≤" y "≥". Es decir, $P(Z < Z_i) = P(Z \leq Z_i)$ y $P(Z > Z_i) = P(Z \geq Z_i)$.

directa consiste en: (1) buscar las probabilidades acumuladas correspondientes a cada puntuación Z y (2) restar la menor de ellas a la mayor:

$$P(Z < 1{,}14) - P(Z < 0{,}25) = F(1{,}14) - F(0{,}25) = 0{,}8729 - 0{,}5987 = 0{,}2742$$

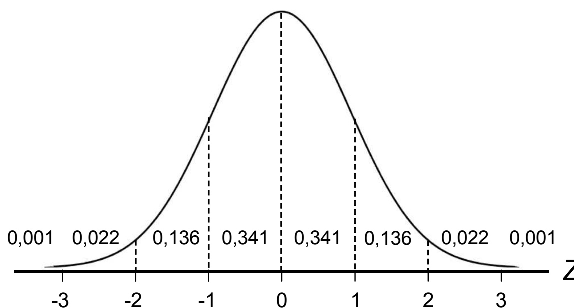
Por supuesto, la tabla también permite conocer cuál es la puntuación Z que acumula una determinada probabilidad. Para saber, por ejemplo, qué puntuación Z acumula una probabilidad de 0,90, se comienza buscando en los valores interiores de la tabla el valor 0,90 (si no existe el valor exacto, se busca el más cercano). Encontramos el valor 0,8997. La puntuación Z que asigna la tabla a una probabilidad acumulada de 0,8997 es 1,28. Por tanto, $Z_{0,90} = 1{,}28$.

Figura 5.4. Probabilidades asociadas a algunas puntuaciones Z en una curva normal



La Figura 5.5 se ha construido con la información que ofrece la tabla de la curva normal. Muestra la proporción de área comprendida entre algunas puntuaciones Z (una, dos y tres desviaciones típicas por encima y por debajo de la media; recuérdese que las puntuaciones Z tienen media 0 y desviación típica 1). Aunque el eje de abscisas de una curva normal admite valores desde $-\infty$ hasta $+\infty$, la figura muestra que el 99,74% de los casos está comprendido entre ±3 puntaciones típicas (o, lo que es lo mismo, entre ±3 desviaciones típicas). Además, en una curva normal tipificada, el 90% del área bajo la curva se encuentra entre ±1,645 puntuaciones típicas; el 95%, entre ±1,96 puntuaciones típicas; el 99%, entre ±2,575 puntuaciones típicas; etc.

Figura 5.5. Proporción de área (probabilidad) entre puntuaciones Z en una curva normal



Por supuesto, las probabilidades de una curva normal también pueden obtenerse mediante un programa informático (sin necesidad, por tanto, de recurrir a la tabla de la curva normal). Ver más adelante, en este mismo capítulo, el apartado *Puntuaciones típicas y curva normal con SPSS*.

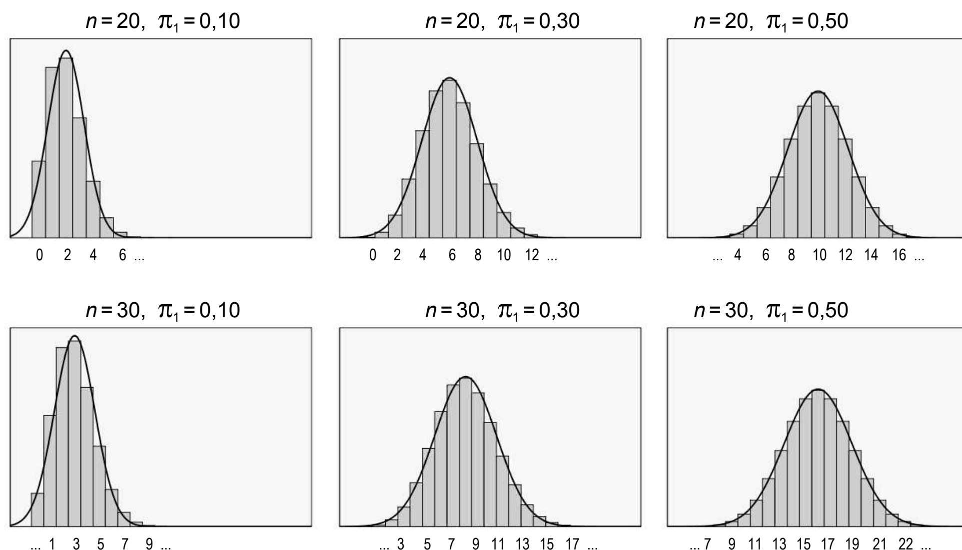
Aproximación de la distribución binomial a la normal

En el Capítulo 3 hemos estudiado la distribución binomial como referente teórico de las probabilidades asociadas al *número de éxitos* observados en un conjunto de n ensayos o réplicas de una variable dicotómica. Y hemos visto que la tabla que ofrece las probabilidades binomiales (Tabla C del Apéndice final) tiene una utilidad algo limitada porque sólo incluye algunos valores de n (número de ensayos) y π_1 (probabilidad de éxito). Para valores de n o π_1 distintos de los que incluye la tabla es necesario utilizar un programa informático (ver Capítulo 3) o la estrategia que se explica en este apartado.

Una de las consecuencias del *teorema del límite central* (ver más arriba) es que la distribución binomial (como otras muchas) se va pareciendo más y más a la distribución normal a medida que el tamaño muestral va aumentando (de hecho, la distribución normal es el límite al que tiende la binomial cuando aumenta n). El grado de parecido es alto incluso con tamaños muestrales relativamente pequeños. Si π_1 toma un valor centrado (próximo a 0,5), tamaños muestrales tan pequeños como 5 permiten obtener ya una buena aproximación; cuanto más se aleja π_1 de 0,5, mayor necesita ser el tamaño muestral para que la aproximación sea buena.

Para ayudar a entender esto, los gráficos de la Figura 5.6 muestran diferentes distribuciones binomiales con sus correspondientes curvas normales superpuestas. Los tres gráficos de la mitad superior se han obtenido con $n=20$ (el valor más grande que ofrece la tabla de la distribución binomial del Apéndice final); los tres de la mitad inferior, con $n=30$. Se ha utilizado $\pi_1 = 0{,}10$ en los dos gráficos de la izquierda, $\pi_1 = 0{,}30$ en los dos del centro y $\pi_1 = 0{,}50$ en los dos de la derecha. La altura de las barras corresponde a las probabilidades binomiales, es decir, a las probabilidades asociadas a 0, 1, 2, 3, ..., n éxitos en n ensayos. La curva normal superpuesta sobre las barras da una idea del grado de parecido existente entre las probabilidades binomiales y las que se obtendrían utilizando la correspondiente curva normal.

Figura 5.6. Grado de parecido entre las probabilidades binomiales y las probabilidades normales con $n=20$ (mitad superior) y $n=30$ (mitad inferior)



Cuando π_1 toma un valor extremo (0,10) la aproximación no es demasiado buena con $n = 20$ (particularmente con los valores más bajos), pero mejora sensiblemente con $n = 30$. Cuando π_1 toma un valor más centrado (0,30) la aproximación empieza a ser muy buena con $n = 20$ (en todo el rango de valores) y mejor todavía con $n = 30$. Cuando π_1 vale 0,50, la aproximación es excelente tanto con $n = 20$ como con $n = 30$. Por supuesto, con tamaños muestrales mayores, el grado de parecido entre ambas distribuciones es todavía mayor.

Justamente este parecido entre ambas distribuciones es el que justifica que las probabilidades normales puedan utilizarse en lugar de las binomiales. Ahora bien, como para conocer las probabilidades normales es necesario utilizar la distribución normal *tipificada*, para poder utilizar las probabilidades normales en lugar de las binomiales es necesario comenzar *tipificando la variable binomial*, es decir, tipificando la variable n_1 = «número de éxitos en n ensayos». Para ello, sabemos que el valor esperado de n_1 es $n\pi_1$ y su varianza $n\pi(1 - \pi_1)$ (ver ecuaciones [3.7] y [3.9]). Por tanto, cualquier valor de n_1 puede transformarse en puntuación Z de la siguiente manera:

$$Z = \frac{n_1 - n\pi_1}{\sqrt{n\,\pi_1(1-\pi_1)}} \qquad [5.9]$$

Ahora bien, si, de acuerdo con el ya mencionado teorema del límite central, n_1 tiende a distribuirse $N\,[np_1,\ (n\pi_1(1-\pi_1))^{1/2}]$ conforme n aumenta, entonces, de acuerdo con [5.7], la transformación Z propuesta en [5.9] tenderá a distribuirse $N(0,\ 1)$.

Para terminar de precisar en qué medida las probabilidades binomiales pueden parecerse a las normales falta aclarar una cuestión. Al hablar de la probabilidad de obtener n_1 éxitos en n ensayos, se está hablando de una variable discreta. Por ejemplo, en 10 ensayos puede haber 0, 1, 2, etc., éxitos, pero no 4,7 éxitos. Por tanto, cuando se utilizan las probabilidades normales en lugar de las binomiales, se está utilizando una distribución *continua* (la normal) en lugar de una distribución *discreta* (la binomial). Para que el grado de parecido entre las probabilidades de ambas distribuciones sea aún mayor se puede intentar hacer que, de alguna manera, los valores discretos se conviertan en continuos. Esto puede conseguirse aplicando lo que se conoce como **corrección por continuidad**, que consiste en considerar que obtener 3 éxitos equivale a obtener éxitos comprendidos entre 2,5 y 3,5; o que obtener menos de 3 éxitos equivale a obtener menos de 2,5 éxitos; o que obtener más de 3 éxitos equivale a obtener más de 3,5 éxitos; etc.

Unos sencillos cálculos pueden ayudar a valorar el grado de parecido existente entre las probabilidades que se obtienen con las distribuciones binomial y normal. Comencemos con el primer gráfico de la Figura 5.6, donde $n = 20$ y $\pi_1 = 0{,}10$ (el peor de los escenarios propuestos). Según la distribución binomial (ver Tabla B del Apéndice final), la probabilidad acumulada asociada a cada número de éxitos vale

$$P(n_1 \leq 0) = F(0) = 0{,}122$$
$$P(n_1 \leq 1) = F(1) = 0{,}392$$
$$P(n_1 \leq 2) = F(2) = 0{,}677$$
$$P(n_1 \leq 3) = F(3) = 0{,}867$$
$$P(n_1 \leq 4) = F(4) = 0{,}957$$
$$\ldots$$

Utilizando la transformación Z y la curva normal con valor esperado $n\pi_1 = 20(0{,}10) = 2$ y desviación típica $[n\pi_1(1-\pi_1)]^{1/2} = [20(0{,}10)(1-0{,}10)]^{½} = 1{,}34$, se obtiene

$$P(n_1 \leq 0) \;=\; P[Z \leq (0{,}5-2)/1{,}34] \;=\; P(Z \leq -1{,}12) \;=\; 0{,}131$$
$$P(n_1 \leq 1) \;=\; P[Z \leq (1{,}5-2)/1{,}34] \;=\; P(Z \leq -0{,}37) \;=\; 0{,}356$$
$$P(n_1 \leq 2) \;=\; P[Z \leq (2{,}5-2)/1{,}34] \;=\; P(Z \leq 0{,}37) \;=\; 0{,}644$$
$$P(n_1 \leq 3) \;=\; P[Z \leq (3{,}5-2)/1{,}34] \;=\; P(Z \leq 1{,}12) \;=\; 0{,}869$$
$$P(n_1 \leq 4) \;=\; P[Z \leq (4{,}5-2)/1{,}34] \;=\; P(Z \leq 1{,}18) \;=\; 0{,}969$$
$$\ldots$$

El parecido entre ambas soluciones es evidente (la diferencia más grande vale $0{,}392 - 0{,}356 = 0{,}036$; es decir, 36 milésimas) a pesar de tratarse del peor de los escenarios propuestos en la Figura 5.6 (pues el valor $\pi_1 = 0{,}10$ está muy alejado de 0,50).

En cualquiera de los restantes escenarios el grado de parecido entre ambas soluciones es todavía mayor. Por ejemplo, la probabilidad de encontrar valores menores o iguales que 1 (que es donde se observa la anterior diferencia de 36 milésimas) utilizando $n = 30$ en lugar de $n = 20$ vale, con la solución binomial[5], 0,184; y con la solución normal, 0,181 (3 milésimas de diferencia). Con $n = 30$ y $\pi_1 = 0{,}10$, la mayor diferencia entre ambas soluciones vale 30 milésimas (se da con la probabilidad de obtener 2 éxitos o menos).

Aunque lo habitual es trabajar con probabilidades *acumuladas*, las probabilidades normales también pueden utilizarse para estimar probabilidades binomiales *individuales*. Consideremos, por ejemplo, la probabilidad de obtener 4 éxitos con $n = 20$ y $\pi_1 = 0{,}30$. La distribución binomial ofrece la siguiente solución: $F(4) - F(3) = 0{,}238 - 0{,}107 = 0{,}131$. La curva normal puede utilizarse para obtener una estimación de ese resultado calculando la probabilidad de obtener valores comprendidos entre 3,5 y 4,5 éxitos (que son los límites que corresponde utilizar de acuerdo con el criterio seguido al aplicar la corrección por continuidad). Así,

$$P(n_1 \leq 4{,}5) \;=\; P[Z \leq (4{,}5-6)/2{,}05] \;=\; P(Z \leq -073) \;=\; 0{,}233$$
$$P(n_1 \leq 3{,}5) \;=\; P[Z \leq (3{,}5-6)/2{,}05] \;=\; P(Z \leq -1{,}22) \;=\; 0{,}111$$

La diferencia entre ambas probabilidades vale 0,122, que es un valor bastante parecido al que ofrece la distribución binomial (9 milésimas de diferencia).

De todo lo anterior cabe concluir que las probabilidades de la curva normal pueden utilizarse para estimar las probabilidades binomiales, lo cual viene a confirmar la generalidad de la curva normal como referente de muchos de los eventos que interesa estudiar.

Puntuaciones típicas y curva normal con SPSS

A diferencia de lo que ocurre con el procedimiento **Frecuencias**, que contiene opciones para describir tanto variables categóricas como cuantitativas continuas (ver Capítulo 3), el procedimiento **Descriptivos** está diseñado únicamente para variables cuantitativas. Incluye algunos

[5] Esta probabilidad puede obtenerse en el SPSS mediante la opción **Calcular** del menú **Transformar**, utilizando, como expresión numérica: CDF.BINOM (1,30,0.1).

estadísticos de tendencia central (media), de dispersión (amplitud total, varianza y desviación típica) y de forma (asimetría, curtosis). Aunque todos estos estadísticos también están incluidos en el procedimiento **Frecuencias**, el procedimiento **Descriptivos** añade en exclusiva una opción importante: la posibilidad de obtener puntuaciones típicas (puntuaciones *Z*). La opción **Guardar valores tipificados como variables** (disponible en el cuadro de diálogo principal) crea en el *Editor de datos* una nueva variable con las puntuaciones típicas correspondientes a cada caso del archivo[6].

Para utilizar el procedimiento **Descriptivos**: (1) seleccionar la opción **Estadísticos descriptivos > Descriptivos** del menú **Analizar** y (2) trasladar una o más variables a la lista **Variables** (la especificación mínima requerida es una variable numérica; la lista de variables del archivo de datos ofrece un listado de todas las variables con formato numérico; las variables con formato de cadena no están disponibles).

Por lo que se refiere a la curva normal, el SPSS incluye varias funciones para trabajar con sus probabilidades. Estas funciones se encuentran en la opción **Calcular** del menú **Transformar**. La función **CDF.NORMAL(*y*,*m*,*s*)** calcula la probabilidad acumulada hasta el valor *y* en una distribución normal con media *m* y desviación típica *s*. Las probabilidades de la Figura 5.3 pueden obtenerse de la siguiente manera:

CDF.NORMAL(−0.82, 0, 1) = 0.2061

1 − CDF.NORMAL(1.14, 0, 1) = CDF.NORMAL(−1.14, 0, 1) = 0.1271

CDF.NORMAL(1.14, 0, 1) − CDF.NORMAL(0.25, 0, 1) = 0.2742

Aunque en estos tres ejemplos se ha utilizado una media igual a 0 y una desviación típica igual a 1, lo cierto es que la función CDF.NORMAL permite elegir cualquier valor tanto para la media como para la desviación típica. Además, para trabajar con la curva normal tipificada, la función **CDF.NORMAL(*y*, *m*, *s*)** puede sustituirse por **CDFNORM(*Z*)**. En esta versión abreviada de la función únicamente hay que indicar la puntuación *Z* cuya probabilidad acumulada se desea conocer.

La función **IDF.NORMAL(*p*, *m*, *s*)** devuelve el valor de la puntuación que acumula una probabilidad *p* en una distribución normal con media *m* y desviación típica *s*. Así, por ejemplo,

IDF.NORMAL(0.2061, 0, 1) = −0.82

1 − IDF.NORMAL(0.1271, 0, 1) = 1.14

Por último, la función **PDF.NORMAL(*y*,*m*,*s*)** permite obtener el valor de la función de densidad, es decir, la densidad que la ecuación [5.5] asigna a *y* en una distribución normal con media *m* y desviación típica *s*. En general, esta función tiene poca utilidad; lo que suele interesar a un analista de datos es conocer la proporción de área que queda por debajo o por encima de cada valor del eje de abscisas.

Debe tenerse en cuenta que el separador decimal que debe utilizarse en la expresiones numéricas del SPSS es el punto (como en una calculadora), no la coma (como se hace al escribir en español).

---

[6] Esta nueva variable recibe, por defecto, el nombre de la variable original más el prefijo *Z*. Si el nombre de variable que resulta de añadir el prefijo *Z* ya existe, entonces el nombre que recibe la nueva variable es *Zsco01*; si durante la misma sesión se tipifica una segunda variable, recibe el nombre de *Zsco02*; y así sucesivamente hasta *Zsco99* (el prefijo *Zsco* proviene de *Z score*).

Apéndice 5

Las distribuciones binomial y normal, ya estudiadas, no son las únicas distribuciones teóricas disponibles. En los próximos capítulos tendremos ocasión de comprobar que, al trabajar con algunos estadísticos, es necesario recurrir a otro tipo de distribuciones. Podríamos decir que, puesto que las distribuciones teóricas de probabilidad son modelos que intentan representar el mundo real y el mundo real es diverso, para poder dar cuenta de diferentes aspectos de la realidad es necesario recurrir a diferentes modelos.

En este apéndice se describen dos de las distribuciones de probabilidad teóricas que, junto con la binomial y la normal, más se utilizan al analizar datos[7]: la distribución χ^2 y la distribución *t*. En ambos casos se trata de distribuciones teóricas (también lo son la binomial y la normal) que no serían objeto de nuestra atención aquí si no fuera porque están estrechamente relacionadas con algunos de los procedimientos estadísticos que estudiaremos.

La distribución χ^2

La distribución χ^2 (letra griega *ji* al cuadrado)[8], propuesta por Pearson en 1900, se obtiene, al igual que el resto de distribuciones teóricas, a partir de una ecuación matemática. No obstante, antes de ofrecer esa ecuación vamos a intentar formarnos una idea algo más intuitiva de esta distribución.

Supongamos que la distribución poblacional de la variable Y es normal con parámetros μ y σ^2. Es decir, supongamos que $Y \sim N(\mu_Y, \sigma_Y)$. Seleccionemos aleatoriamente una observación de esa población y calculemos su puntuación típica Z elevada al cuadrado:

$$Z^2 = \frac{(Y - \mu_Y)^2}{\sigma_Y^2} \qquad [5.10]$$

Llamemos χ_1^2 a esta transformación. Es decir, hagamos

$$\chi_1^2 = Z^2 \qquad [5.11]$$

(enseguida explicaremos el significado del subíndice 1). Sigamos seleccionando una a una observaciones de la población de Y hasta obtener los valores $\chi_1^2 = Z^2$ correspondientes a todos los posibles valores Y. ¿Qué podemos decir de estas nuevas puntuaciones χ_1^2? Sabemos que, antes de ser elevadas al cuadrado, las puntuaciones Z toman valores entre $-\infty$ y $+\infty$. Sin embargo, después de elevadas al cuadrado, todas ellas se vuelven positivas: las nuevas puntuaciones χ_1^2 toman valores entre 0 y $+\infty$.

Esta transformación produce un cambio sustancial en la forma de la distribución. Las curvas de la Figura 5.7 pueden ayudar a entender este cambio. La curva de la izquierda indica que, en una distribución normal tipificada, el 68,26% de las puntuaciones Z se encuentra dentro del rango (−1, 1); al elevar al cuadrado estas puntuaciones (curva de la derecha), todas ellas pasan a formar parte del rango (0, 1). Del mismo modo, al elevar al cuadrado el 15,87% de las puntuaciones Z menores que −1 y el 15,87% de las mayores que 1, todas ellas (31,74%) pasan a tomar valores mayores que 1.

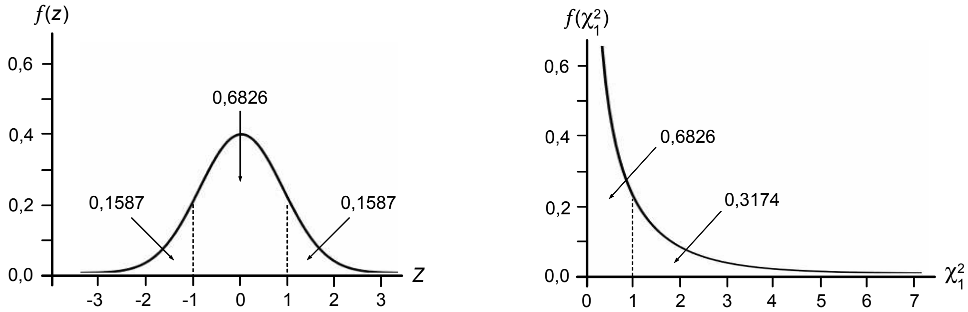
---

[7] Todas estas distribuciones están relacionadas. La distribución normal representa el límite al que tiende la binomial conforme *n* va aumentando. Y las distribuciones χ^2 y *t* se derivan, ambas, de la normal. Todas ellas pertenecen a una familia de distribuciones llamada *exponencial*.

[8] Lo que aquí estamos llamando "*ji*-cuadrado" (χ^2) también puede encontrarse en la literatura estadística en español con el nombre "*chi*-cuadrado", anglicismo innecesario e incorrecto que, sin embargo, nos veremos obligados a utilizar por ser el que aparece en el SPSS.

La consecuencia más destacable de esta transformación es que, aunque la distribución de las puntuaciones Z (curva de la izquierda) es simétrica, la distribución de las nuevas puntuaciones χ_1^2 (curva de la derecha) es muy asimétrica[9]. Esta distribución recibe el nombre de *ji-cuadrado con 1 grado de libertad* y se obtiene, simplemente, elevando al cuadrado las puntuaciones Z de una distribución normal.

Figura 5.7. Distribución normal tipificada (izquierda) y distribución χ^2 con 1 grado de libertad (derecha)



Supongamos ahora que en lugar de seleccionar de la población de Y valores uno a uno, los seleccionamos de dos en dos (Y_1 e Y_2) y, al igual que en [5.11], calculamos Z_1^2 a partir de Y_1 y Z_2^2 a partir de Y_2. Llamemos χ_2^2 a la suma de estas dos puntuaciones típicas elevadas al cuadrado. Es decir, hagamos

$$\chi_2^2 = Z_1^2 + Z_2^2 \qquad [5.12]$$

Repitiendo el proceso con todas las posibles muestras de tamaño 2 podemos obtener la distribución de las nuevas puntuaciones χ_2^2. Todas ellas son, al igual que las puntuaciones χ_1^2, positivas. Pero la distribución de χ_2^2 es menos asimétrica que la de χ_1^2, pues la probabilidad de encontrar valores comprendidos entre 0 y 1 se ha reducido sustancialmente. Ahora, al sumar dos puntuaciones Z^2 que se mueven en el rango (0, 1), se obtienen puntuaciones dentro del rango (0, 2). Del mismo modo, al sumar dos puntuaciones Z^2 con valores mayores que 1 se obtienen puntuaciones mayores que 2.

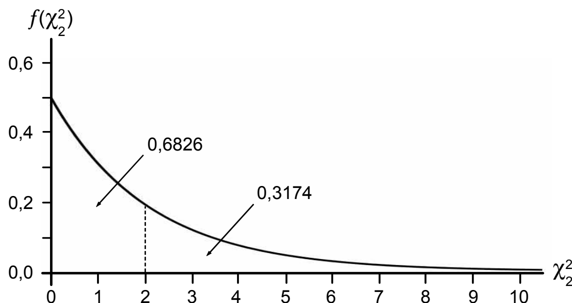
La distribución resultante de esta transformación adopta la forma que muestra la Figura 5.8. Recibe el nombre de distribución *ji-cuadrado con 2 grados de libertad* y se obtiene, simplemente, elevando al cuadrado y sumando dos puntuaciones típicas de una distribución normal.

La diferencia entre las variables χ_1^2 y χ_2^2 está únicamente en el número de puntuaciones Z^2 que contribuyen al resultado. En χ_1^2 sólo hay una puntuación (Y) que puede variar libremente (puede tomar cualquiera de sus posibles valores) y aportar información nueva; por tanto, sólo hay un dato que contribuye a generar el resultado χ_1^2. En χ_2^2 hay dos puntuaciones independientes (Y_1 e Y_2) que pueden variar libremente y generar información nueva; por tanto, hay dos datos independientes que contribuyen a generar el resultado χ_2^2. Ésta es la idea que se recoge en el concepto de *grados de libertad* en una distribución ji-cuadrado: número de unidades (observaciones) independientes que, variando libremente dentro de su dominio, aportan nueva. Los subíndices que acompañan a χ^2 indican el número de grados de libertad. Por tanto, χ_1^2 es una variable ji-cuadrado con 1 grado de libertad; y χ_2^2 es una variable ji-cuadrado con 2 grados de libertad.

---

9 Para entender bien esta curva hay que tener en cuenta que una *densidad* no es una *probabilidad*. Una densidad puede tomar valores mayores que uno. En una curva normal tipificada (Figura 5.7, izquierda), todas las densidades son menores de 0,4. Sin embargo, en la distribución ji-cuadrado con 1 grado de libertad (Figura 5.7, derecha), los valores ji-cuadrado menores de 0,14 tienen densidades mayores que 1. Ésta es la razón por la cual la curva del dibujo no toca el eje vertical: habría que dibujar una curva desproporcionadamente alta y estrecha para poder representar el punto de aproximación. Por ejemplo, un valor ji-cuadrado de 0,01 (eje horizontal) tiene una densidad de 3,97 (eje vertical); un valor de 0,001 tiene una densidad de 12,61; un valor de 0,0001 tiene una densidad de 39,89; un valor de 0,00001 tiene una densidad de 126,16; etc.

Figura 5.8. Distribución χ^2 con 2 grados de libertad



Supongamos, por último, que de la población normal de puntuaciones *Y* seleccionamos *k* observaciones independientes y, con cada una de ellas, efectuamos la transformación

$$Z_i^2 = \frac{(Y_i - \mu_Y)^2}{\sigma_Y^2} \qquad [5.13]$$

(con $i = 1, 2, ..., n$). Llamemos χ^2_n a la suma de estas *n* puntuaciones Z_i^2. Es decir, hagamos:

$$\chi^2_n = \sum_{i=1}^{n} Z_i^2 \qquad [5.14]$$

Repitiendo el proceso con todas las posibles muestras de *n* observaciones se puede obtener la distribución de las nuevas puntuaciones χ^2_n. Ahora bien, para conocer esa distribución no es necesario seleccionar todas las posibles muestras de tamaño *n*. Sabemos que, dado un valor de *n*, la función que asigna una densidad concreta a cada posible valor χ^2 viene dada por

$$f(\chi^2 \,|\, n) = \frac{1}{2^{n/2}\,\Gamma(n/2)} \left[\chi^2\right]^{(n/2)-1} \left[e\right]^{-\chi^2/2} \qquad (\text{con } Y \geq 0 \text{ y } n \geq 1) \qquad [5.15]$$

(el término $\Gamma(n/2)$ del denominador es una función denominada *gamma*; ver Winer, Brown y Michels, 1991, págs. 852-856). Aunque es seguro que esta ecuación puede resultar bastante desagradable para muchos lectores, debe tenerse en cuenta que no será necesario trabajar con ella (existen tablas y programas de ordenador que facilitan el trabajo). Basta con conocer algunas de sus características.

La función [5.9] únicamente tiene un parámetro, *n*, el cual ya sabemos que recibe el nombre de *grados de libertad* (recordemos que la distribución normal tiene dos parámetros: μ y σ; y la distribución binomial otros dos: *n* y π). Esto significa que la forma de la distribución de χ^2_n únicamente cambia cuando cambia *n*. El resto de elementos de la ecuación, exceptuando χ^2_n, son constantes.

Puesto que cada valor de *n* genera una distribución distinta, no existe una única distribución χ^2, sino una familia de distribuciones χ^2: todas ellas son asimétricas positivas (particularmente con pocos grados de libertad), pero el grado de asimetría se va reduciendo conforme va aumentando *n*. En las distribuciones que muestra la Figura 5.9 se aprecia claramente que las curvas se van ensanchando, aplastando y perdiendo asimetría conforme van aumentando los grados de libertad[10].

[10] Si se desea comparar las curvas de la Figura 5.9 con las curvas de las Figuras 5.7 y 5.8 para obtener una idea lo más exacta posible sobre la forma de la distribución χ^2, debe tenerse en cuenta que las escalas del eje horizontal y del vertical de esas curvas no son las mismas. Las curvas de las Figuras 5.7 y 5.8 están dibujadas utilizando la misma anchura en las unidades del eje horizontal y la misma altura en las unidades del eje vertical; por tanto, sus formas son comparables. Pero en las curvas de la Figura 5.9 se ha cambiado la escala de ambos ejes. Las unidades del eje vertical se han hecho más grandes y las del eje vertical más pequeñas para evitar tener que dibujar curvas demasiado bajas y demasiado anchas.

Dadas las características de una variable χ^2, es fácil comprobar que su centro o valor esperado es precisamente *n* (los grados de libertad) y que su varianza es el doble de *n*:

$$E(\chi^2_n) = \mu_{\chi^2_n} = \sum_i^n E(\chi^2_i) = n$$
$$V(\chi^2_n) = \sigma^2_{\chi^2_n} = \sum_i^n V(\chi^2_i) = 2n \qquad [5.16]$$

De la ecuación [5.14] se deduce una interesante propiedad de las variables ji-cuadrado: la suma de *k* variables ji-cuadrado es a su vez una variable ji-cuadrado cuyos grados de libertad son el resultado de sumar los de las *k* variables sumadas; es decir,

$$\chi^2_{n_1} + \chi^2_{n_2} + \cdots + \chi^2_{n_k} = \chi^2_{n_1 + n_2 + \cdots + n_k} \qquad [5.17]$$

Figura 5.9. Distribuciones χ^2 con 5, 10, 15, 20 y 30 grados de libertad

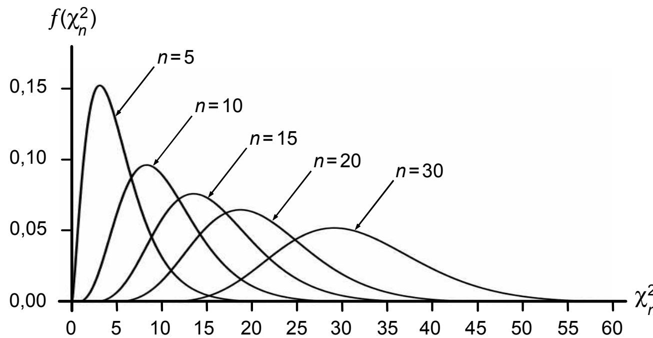


Tabla de la distribución χ^2

Por lo general, el trabajo con la distribución χ^2 se limita a calcular la proporción de área bajo la curva asociada a los valores del eje horizontal o, alternativamente, el valor concreto del eje horizontal asociado a una determinada proporción de área bajo la curva. Al realizar esta tarea, se asume que el área bajo la curva vale 1 y, por tanto, que hablar de *proporción de área* es equivalente a hablar de *probabilidad*.

La Tabla D del Apéndice final ofrece los valores χ^2 correspondientes a algunos cuantiles. La primera columna de la tabla recoge los grados de libertad (*gl*); por tanto, cada fila de la tabla corresponde a una curva χ^2 distinta. Los valores del interior de la tabla son los valores χ^2 del eje horizontal de la curva. Para leer correctamente la tabla hay que tener en cuenta, tal como se indica junto al título de la tabla, que los valores de las cabeceras de las columnas se refieren a probabilidades *acumulada*s, es decir, a la proporción de área que queda por debajo de cada valor χ^2.

En la tabla, el valor χ^2 correspondiente al cruce entre la segunda fila ($gl = 2$) y la séptima columna ($P = 0{,}95$) vale 5,99. Esto significa que, en la distribución χ^2 con 2 grados de libertad, el valor 5,99 acumula (o sea, deja por debajo o a la izquierda) una proporción de área de tamaño 0,95. Es decir,

$$P(\chi^2_2 < 5{,}99) = F(5{,}99) = 0{,}95$$

Ahora bien, si por debajo del valor 5,99 queda una proporción de área de tamaño 0,95, entonces la probabilidad de encontrar valores χ^2_2 menores que 5,99 vale 0,95 (esto es lo que se quiere indicar al decir que *proporción de área* es equivalente a *probabilidad*). Para representar este resultado utilizamos la siguiente expresión:

$$\chi^2_{2;\,0{,}95} = 5{,}99$$

Los dos subíndices de χ^2 suelen colocarse separados por un punto y coma para evitar confusiones con la coma decimal; el primero de ellos se refiere a los grados de libertad (*gl*); el segundo, a la proporción de área que deja a la izquierda o que acumula cada valor χ^2 (probabilidad acumulada). La tabla únicamente ofrece algunas distribuciones (hasta 30 *gl*) y algunos cuantiles de esas distribuciones. Para conocer otros valores puede utilizarse un programa informático como el SPSS.

La distribución χ^2 con SPSS

El SPSS incluye varias funciones relacionadas con la distribución χ^2 (la cual recibe el nombre de *chi-cuadrado*; del vocablo inglés *chi-square*). Todas ellas se encuentran en la opción **Calcular** del menú **Transformar**.

La función **CDF.CHISQ (*x*, *n*)** calcula la probabilidad acumulada hasta el valor x en una distribución χ^2 con n grados de libertad. Por tanto, para calcular el cuantil correspondiente al valor 5,99 en una distribución χ^2 con 2 grados de libertad, haremos

CDF.CHISQ (5.99, 2) = 0.95

Restando de 1 la expresión anterior se obtiene la probabilidad que queda por encima de 5.99. Y restando las expresiones correspondientes a dos valores ji-cuadrado se obtiene la probabilidad comprendida entre ambos.

La función **IDF.CHISQ (*p*, *n*)** devuelve el valor de la puntuación que acumula una probabilidad p en una distribución χ^2 con n grados de libertad. Así, la siguiente expresión permite conocer el valor que corresponde al cuantil 95 en una distribución con 2 grados de libertad:

IDF.CHISQ (0.95, 2) = 5.99

Por último, la función **PDF.CHISQ (*x*,*n*)** permite obtener el valor de la función de densidad, es decir, el valor que la ecuación 5.15 asigna a x en una distribución χ^2 con n grados de libertad. No debe olvidarse que el separador decimal que debe utilizarse en la expresiones numéricas del SPSS es el punto (como en una calculadora), no la coma (como se hace al escribir en español).

La distribución *t*

La distribución t, propuesta por W. S. Gosset en 1908, también es conocida como distribución t de *Student* (seudónimo con el que Gosset publicó sus trabajos). La función de densidad de esta distribución viene dada por

$$f(t) = \frac{\Gamma\,[(n+1)/2]}{\Gamma\,[n/2]\,\sqrt{n\pi}}\,(1+t^2/n)^{-(n+1)/2} \qquad [5.18]$$

La forma de esta función de densidad únicamente cambia cuando cambia el valor de n (para un determinado valor de n, todos los elementos de la ecuación son constantes excepto los valores t cuya densidad se desea conocer). El único parámetro de la función es n. Y, al igual que en la distribución ji-cuadrado, n representa los *grados de libertad*. Esto significa que no existe una única distribución t sino una familia de distribuciones, todas las cuales se derivan de la función [5.18] asignando valores a t; la forma exacta de esas distribuciones depende del parámetro n (los grados de libertad).

Una rápida inspección de la función [5.18] permite comprobar que a los valores t que tienen el mismo valor absoluto se les asigna la misma densidad (pues el valor t está elevado al cuadrado). Además, la densidad más alta corresponde al valor $t = 0$ (pues el resultado de elevar un número a una potencia negativa es tanto mayor cuanto más próximo a cero está el número) y va disminuyendo conforme t se va alejando de cero. De estas dos consideraciones se desprende que la curva resultante de [5.18] es

simétrica respecto de su centro (que vale cero), unimodal (la densidad más alta corresponde al valor central) y con forma de campana (las densidades van disminuyendo conforme se alejan del centro). Y sabemos que, aunque su centro vale cero, su varianza está relacionada con los grados de libertad:

$$E(t) = \mu_t = 0$$
$$V(t) = \sigma_t^2 = n/(n-2) \qquad \text{(para } n > 2) \qquad [5.19]$$

Es decir, todo en la distribución t es exactamente igual que en la curva normal tipificada excepto la varianza (1 frente a $n(n-1)/2$; ver, por ejemplo, Hogg y Craig, 1970, págs. 166-167).

La fórmula de la varianza de t revela algo muy importante: su valor es tanto menor cuanto mayor es el tamaño muestral (pues la diferencia relativa entre n y $n-2$ es tanto menor cuanto mayor es n). De hecho, el valor de la varianza se va aproximando a 1 conforme va aumentando n; lo cual implica que la curva t se va pareciendo más y más a la normal a medida que va aumentando n.

La Figura 5.10 muestra una curva t con 2 grados de libertad (t_2), otra con 5 grados de libertad (t_5) y una distribución normal tipificada $N(0, 1)$. Con pocos grados de libertad, la distribución t es algo más ancha y aplastada que la normal tipificada (respecto de la normal, la proporción de área es mayor en las colas y menor en el centro). Pero, conforme aumentan los grados de libertad, la forma de la distribución t se va aproximando rápidamente a la de la distribución normal. A partir de 10 grados de libertad (ver Figura 5.11), la distribución t apenas se distingue de la normal tipificada.

Figura 5.10. Distribución normal tipificada junto con dos distribuciones *t* (2 y 5 grados de libertad)

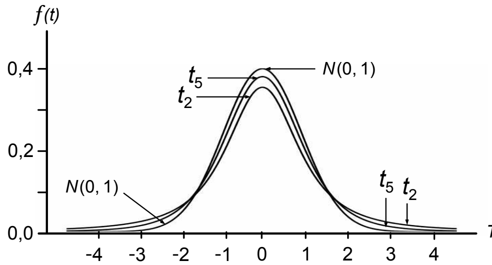
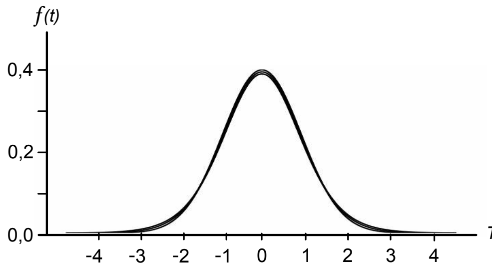


Figura 5.11. Distribución normal tipificada junto con dos distribuciones *t* (10 y 20 grados de libertad)



Lo interesante de la función [5.18] es que existen algunas transformaciones muy utilizadas al analizar datos que se ajustan a la distribución teórica que se deriva de ella. Por ejemplo, si una variable Z distribuida $N(0, 1)$ y una variable χ_n^2 distribuida según *ji*-cuadrado con n grados de libertad, ambas independientes, se combinan de la siguiente manera:

$$T = \frac{Z}{\sqrt{\chi_n^2 / n}} \qquad [5.20]$$

se obtiene una variable T cuya función de densidad es precisamente [5.18]. De hecho, con cierta frecuencia se utiliza la transformación [5.20] para definir la distribución t. Más adelante se discuten algunos detalles relacionados con esta transformación (ver Apéndice 9).

Tabla de la distribución *t*

Al igual que con el resto de distribuciones estudiadas, el trabajo con la distribución t suele limitarse a calcular la proporción de área bajo la curva asociada a diferentes valores del eje horizontal o, alternativamente, a encontrar el valor concreto del eje horizontal asociado a una determinada proporción de área bajo la curva. Para esta tarea, seguimos asumiendo que toda el área bajo la curva vale 1 y, por tanto, que hablar de *proporción de área* es equivalente a hablar de *probabilidad*.

La Tabla E del Apéndice final ofrece algunos cuantiles de la distribución t. La primera columna de la tabla contiene los grados de libertad (*gl*); por tanto, cada fila de la tabla recoge una curva t distinta (claro está, sólo algunos cuantiles de cada curva t). Los valores del interior de la tabla son los valores t del eje horizontal.

Para leer correctamente la tabla hay que tener en cuenta, tal como se indica en el título de la tabla, que las cabeceras de las columnas son probabilidades *acumuladas*: las proporciones de área que quedan por debajo de los correspondientes valores del eje horizontal.

Consultando la tabla vemos que el valor correspondiente al cruce entre la quinta fila ($gl = 5$) y la sexta columna ($p = 0{,}95$) vale 2,015. Este resultado indica que, en la distribución t con 5 grados de libertad, el valor 2,015 acumula (deja por debajo de sí o a la izquierda) una proporción de área cuyo tamaño es 0,95:

$$P(t_5 < 2{,}015) \;=\; F(2{,}015) \;=\; 0{,}95$$

Por tanto, si por debajo del valor 2,015 queda una *proporción de área* de tamaño 0,95, entonces la *probabilidad* de encontrar, en esa distribución, valores t menores que 2,015 vale 0,95. Para representar este resultado utilizamos la expresión

$$t_{5;\,0{,}95} \;=\; 2{,}015$$

El primer subíndice de t se refiere a los grados de libertad (*gl*); el segundo, a la proporción de área que deja a la izquierda ese valor t (probabilidad acumulada).

La tabla únicamente ofrece algunos valores de la distribución; en concreto los valores de ambas colas: del percentil 0,1 al 10 y del 90 al 99,9. Como la distribución es simétrica, se verifica

$$t_{gl;\,p} \;=\; -t_{gl;\,1-p}$$

Así, por ejemplo, el valor que acumula una probabilidad de 0,05 en la distribución t con 5 grados de libertad es, cambiado de signo, el mismo que acumula una probabilidad de 0,95:

$$t_{5;\,0{,}05} \;=\; -t_{5;\,0{,}95} \;=\; -2{,}015$$

La distribución *t* con SPSS

La opción **Calcular** del menú **Transformar** incluye algunas funciones relacionadas con la curva t. La función **CDF.T (*t*, *n*)** calcula la probabilidad acumulada hasta el valor t en una distribución con n grados de libertad. Por tanto, para calcular el cuantil correspondiente al valor 2,015 en una distribución t con 5 grados de libertad, haremos

CDF.T(2.015, 5) = 0.95

Restando de 1 la expresión anterior se obtiene la probabilidad que queda por encima de 2.015:

1 – CDF.T (2.015, 5) = CDF.T (–2.015, 5) = 0.05

Y restando las expresiones correspondientes a dos valores *t* se obtiene la probabilidad comprendida entre ambos:

CDF.T (2.015, 5) – CDF.T (–2.015, 5) = 0.90

La función **IDF.T (*p*, *n*)** devuelve el valor *t* que acumula una probabilidad *p* en una distribución con *n* grados de libertad. Así, la siguiente expresión permite conocer el valor *t* que corresponde al cuantil 95 en una distribución con 5 grados de libertad:

IDF.T(0.95, 5) = 2.015

Por último, la función **PDF.T (*t*, *n*)** permite obtener el valor de la función de densidad, es decir, el valor que la ecuación [5.18] asigna a *t* en una distribución con *n* grados de libertad.

Ejercicios

5.1. La siguiente tabla recoge las puntuaciones obtenidas al aplicar una prueba de selección (Y) a los 20 candidatos a un puesto de trabajo (las puntuaciones están ordenadas de menor a mayor):

Candidatos	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Y	2	7	10	11	13	14	15	15	16	16	17	17	17	18	18	18	18	19	19	20

En el capítulo anterior ya hemos calculado la media (= 15) y la desviación típica (= 4,51) de estas puntuaciones.

a. ¿Qué puntuación típica corresponde al quinto candidato?

b. En una curva normal, una puntuación típica como la obtenida en el apartado anterior acumula una probabilidad de 0,33. ¿Ocurre eso mismo en la distribución de las puntuaciones de los 20 candidatos?

c. Las puntuaciones originales de la prueba de selección oscilan entre 0 y 20 puntos ¿Qué puede hacerse con estas puntuaciones para que tomen valores entre 0 y 100.

d. ¿Qué tipo de transformación puede aplicarse a las puntuaciones originales para que queden expresadas en una escala con media 5 y desviación típica 2?

e. ¿Cambian las puntuaciones típicas originales al aplicar la transformación del apartado anterior?

f. ¿Es buena idea aplicar a las puntuaciones originales (que varían entre 0 y 20 puntos) la transformación del apartado *d*?

5.2. El responsable de seleccionar entre los aspirantes a un puesto de trabajo, además de la prueba de selección del ejercicio anterior, con media 15 y desviación típica 4,51, ha utilizado otra prueba en la que los sujetos han obtenido una media de 8 y una desviación típica de 3,24.

a. ¿Cómo se pueden combinar las puntuaciones de ambas pruebas si se quiere obtener para cada sujeto una puntuación final en la que ambas pruebas tengan el mismo peso?

b. Si se considera que la primera prueba es el doble de importante que la segunda, ¿cómo se pueden combinar las puntuaciones de ambas pruebas para obtener una puntuación final en la que la primera prueba tenga el doble de peso que la segunda?

c. Un sujeto ha obtenido una puntuación de 14 en la primera prueba y una puntuación de 9 en la segunda; otro sujeto ha obtenido las puntuaciones 17 y 6. Combinando ambas puntuaciones de tal modo que ambas tengan el mismo peso, ¿cuál de los dos sujetos tendrá una puntuación final mayor?

d. Seguimos con los dos sujetos del apartado anterior. Combinando sus puntuaciones de tal modo que la de la primera prueba tenga el doble de peso que la de la segunda, ¿cuál de los dos sujetos tendrá una puntuación final mayor?

e. ¿Qué debe hacer nuestro responsable de selección para que las puntuaciones finales obtenidas en cualquiera de los apartados anteriores queden expresadas en una escala con media 100 y desviación típica 20?

5.3. El siguiente conjunto de puntuaciones procede de una muestra de 10 pacientes a los que se ha administrado la escala de depresión de Hamilton antes (Y_{antes}) y después ($Y_{\text{después}}$) de recibir un tratamiento antidepresivo:

Pacientes	1	2	3	4	5	6	7	8	9	10	$\bar{Y}_j$	S_{Y_j}
Y_{antes}	9	8	21	14	3	17	14	33	22	19	18	8,50
$Y_{\text{después}}$	10	2	11	6	3	10	3	15	8	12	8	4,37

a. Si la puntuación directa de un paciente cualquiera es mayor en el momento *antes* que en el momento *después*, ¿también su puntuación típica será mayor en el momento *antes* que en el momento *después*? Después de razonar la respuesta, calcular las puntuaciones típicas del décimo paciente en ambos momentos.

b. Si un paciente obtiene un puntuación típica mayor que otro en el momento *antes*, ¿también el percentil del primero será mayor que el del segundo?

c. Si la puntuación directa de un paciente en el momento *antes* es mayor que su puntuación directa en el momento *después*, ¿también tendrá un percentil mayor en el momento *antes* que en el momento *después*?

d. Una forma habitual de valorar el cambio que experimentan los sujetos tras el tratamiento consiste en restar las puntuaciones de ambos momentos: $Y_{\text{antes}} - Y_{\text{después}}$. Se obtienen así unas nuevas puntuaciones (D_Y = diferencias) cuya magnitud y signo informan del cambio experimentado. Si se tipifican estas diferencias, ¿las puntuaciones típicas resultantes siguen informando del cambio experimentado?

e. Si, en lugar de tipificar las diferencias del apartado anterior, primero se calculan las puntuaciones típicas de ambos momentos (Z_{antes} y $Z_{\text{después}}$) y luego se restan ($D_Z = Z_{\text{antes}} - Z_{\text{después}}$), ¿se sigue obteniendo información sobre el cambio experimentado entre ambos momentos?

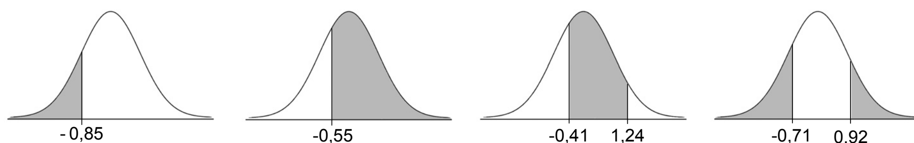
f. Si se comparan las medias de los momentos *antes* y *después* se obtiene una diferencia de 16 − 8 = 8 puntos, lo cual indica que las puntuaciones han bajado, en promedio, 8 puntos. Si se tipifican primero las puntuaciones de ambos momentos y después se obtiene la diferencia entre las medias, ¿se llega a una conclusión equivalente?

5.4. Conocemos algunas puntuaciones de una muestra de 5 sujetos en una variable cualquiera *Y*. Sabemos que la media de los 5 sujetos vale 15 y la desviación típica 4. Y conocemos la puntuación directa de uno de ellos, la puntuación diferencial de otro, la puntuación típica de otro y la puntuación transformada $X = 50 + 10\,Z_Y$ de otro. Completar la tabla.

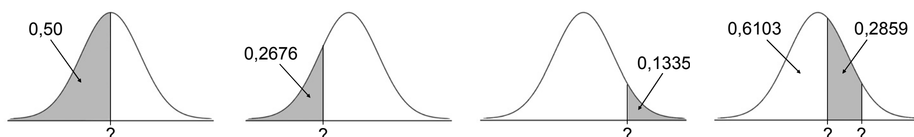
Sujetos	*Y*	*y*	Z_Y	*X*
1	11	( )	( )	( )
2	( )	0	( )	( )
3	( )	( )	0,5	( )
4	( )	( )	( )	45
5	( )	( )	( )	( )

5.5. El objetivo de este ejercicio es el de familiarizarse con la tabla de la curva normal tipificada (Tabla C del Apéndice final). En la mayoría de los ejercicios en los que interviene la curva normal es necesario recurrir a las probabilidades de la curva normal tipificada. Vamos a comenzar utilizando dibujos de la curva.

a. Averiguar qué proporción de área corresponde a cada una de las zonas sombreadas en las siguientes distribuciones:



b. Averiguar qué puntuaciones típicas corresponden a las áreas marcadas en las siguientes distribuciones:



5.6. Seguimos intentando familiarizarnos con el uso de la tabla de la curva normal tipificada, pero ahora sin la ayuda de dibujos. Averiguar qué probabilidad corresponde a cada una de las siguientes expresiones:

a. $P(Z < 0)$
b. $P(Z < -0{,}50)$
c. $P(Z < -1{,}82)$
d. $P(Z < 0{,}50)$
e. $P(Z < 1{,}82)$
f. $P(Z > 0)$
g. $P(Z > 0{,}50)$
h. $P(Z > 1{,}25)$
i. $P(Z > -0{,}50)$
j. $P(Z > -1{,}25)$
k. $P(0 < Z < 0{,}50)$
l. $P(0 < Z < 2{,}15)$
m. $P(0{,}50 < Z < 2{,}15)$
n. $P(1{,}22 < Z < 2{,}15)$
ñ. $P(-0{,}50 < Z < 0)$
o. $P(-1{,}60 < Z < 0)$
p. $P(-1{,}60 < Z < -0{,}50)$
q. $P(-2{,}15 < Z < -1{,}60)$
r. $P(-0{,}50 < Z < 0{,}50)$
s. $P(-1{,}82 < Z < 0{,}50)$
t. $P(-0{,}50 < Z < 1{,}82)$
u. $P(-1{,}60 < Z < 1{,}82)$
v. $P(Z < -1{,}82) + P(Z > 0)$
w. $P(Z < -1{,}82) + P(Z > 1{,}25)$
x. $P(Z < 0) + P(Z > 0{,}50)$
y. $P(Z < 0{,}95) + P(Z > 2{,}15)$
z. $P(Z < 0) + P(Z > 0)$

5.7. Para terminar con este intento de familiarizarnos con el uso de la tabla de la curva normal tipificada, averiguar qué valor toma cada una de las siguientes puntuaciones típicas:

a. $Z_{0,0013}$	*f*. $Z_{0,1003}$	*k*. $Z_{0,9495}$
b. $Z_{0,0052}$	*g*. $Z_{0,2514}$	*l*. $Z_{0,9750}$
c. $Z_{0,0099}$	*h*. $Z_{0,5000}$	*m*. $Z_{0,9901}$
d. $Z_{0,0250}$	*i*. $Z_{0,7486}$	*n*. $Z_{0,9948}$
e. $Z_{0,0505}$	*j*. $Z_{0,8997}$	*ñ*. $Z_{0,9987}$

5.8. Si las puntuaciones directas de cinco sujetos se transforman en puntuaciones típicas, ¿entra dentro de lo posible que esas puntuaciones típicas sean −2, −1, 0, 1 y 2? Justificar la respuesta.

5.9. Las puntuaciones (CI) que ofrece el WAIS (Wechsler Adult Intelligence Scale) se distribuyen normalmente y están transformadas para que la media resultante valga 100 y la desviación típica 15. Esto supuesto:

a. ¿Qué puntuación típica corresponde a un sujeto que obtiene una puntuación directa (es decir, un CI) de 130?

b. ¿Qué porcentaje de sujetos queda por encima de ese sujeto?

c. ¿Entre qué puntuaciones (CI) se encuentra el 95% de los sujetos?

d. ¿Qué puntuación correspondería al sujeto que ha obtenido un CI de 130 en una escala con media 5 y desviación típica 1,5?

e. Supongamos que queremos utilizar las puntuaciones de la escala WAIS para clasificar a los sujetos en cinco grupos de inteligencia: muy baja, baja, media, alta, muy alta. El criterio para formar estos grupos es que los grupos extremos contengan el 5% de los casos y el grupo central el 50%. ¿Qué valores (CI) deben utilizarse como puntos de corte?

5.10. El tiempo necesario para completar un examen académico (Y) se distribuye normalmente con una media de 120 minutos y una desviación típica de 20 minutos.

a. ¿Qué porcentaje de estudiantes cabe esperar que haya completado el examen al cabo de dos horas?

b. ¿Qué porcentaje de estudiantes cabe esperar que haya completado el examen al cabo de 100 minutos?

c. ¿Cuánto tiempo debe durar el examen para que le dé tiempo a terminarlo al 85% de los estudiantes?

5.11. En una curva normal con media y desviación típicas desconocidas, ¿cuál es el porcentaje de puntuaciones que se aleja de la media más de dos desviaciones típicas?

5.12. Tenemos un conjunto de 1.000 puntuaciones que se distribuyen normalmente. Sabemos que por debajo de la puntuación $Y_1 = 147$ hay 63 puntuaciones y que por encima de la puntuación $Y_2 = 541$ hay 8 puntuaciones. ¿Cuánto valen la media y la desviación típica de esas 1.000 puntuaciones?

5.13. Consideremos dos distribuciones normales distintas (Y_1, Y_2), ambas con medias y desviaciones típicas desconocidas. Al seleccionar al azar dos puntuaciones, una de cada distribución:

a. ¿Cuál es la probabilidad de que ambas se encuentren entre los percentiles 25 y 75 de sus respectivas distribuciones?

b. ¿Cuál es la probabilidad de que ambas sean mayores que sus respectivas medias?

c. ¿Cuál es la probabilidad de que ambas estén situadas a más de dos desviaciones típicas por encima de su respectiva media?

5.14. En una distribución normal cuyo percentil 90 vale 40, las puntuaciones directas a las que corresponden las puntuaciones típicas -1 y 1 están separadas 20 puntos. Calcular la media y la desviación típica de la distribución.

5.15. Al encuestar a una muestra aleatoria de 50 personas, cuál es la probabilidad de que más de la mitad se manifieste en contra de la eutanasia si, como se sabe, en la población de donde procede esa muestra, el 40% de las personas está en contra de la eutanasia.

5.16. Una prueba de rendimiento está formada por 30 preguntas, cada una de las cuales consta de 2 alternativas de respuesta de las que sólo una es correcta. Teniendo en cuenta que la respuesta que se da a cada pregunta es independiente de la respuesta que da a las demás:

a. Si un sujeto responde al azar a las 30 preguntas del examen, ¿cuál es el valor esperado de la variable n_1 = «número de aciertos»? ¿Y la varianza?

b. ¿Cuál es la probabilidad de acertar, por azar, más de 20 preguntas?

5.17. Con un tratamiento convencional de la adicción al alcohol se viene obteniendo un 40% de resultados positivos. Un nuevo tratamiento en fase experimental ha conseguido 25 recuperaciones al aplicarlo a 40 pacientes con adicción al alcohol. ¿Qué sugiere este resultado?

5.18. Un examen de la asignatura *Análisis de Datos* consta de 30 preguntas con 5 alternativas de respuesta de las que sólo una es correcta. Para seleccionar un punto de corte con el que aprobar a los estudiantes se ha utilizado un criterio basado en el número de aciertos por azar. En concreto, se tiene intención de aprobar a los estudiantes que tengan más aciertos que los que cabe esperar simplemente respondiendo al azar. Para ello se ha decidido colocar el punto de corte allí donde obtener por azar un número de aciertos igual o mayor que ese punto tenga una probabilidad menor que 0,05. ¿Dónde habrá que colocar el punto de corte?

Soluciones

5.1. *a*. $Z_{(13)} = (13-15)/\,4{,}51 = -0{,}44$ (la media y la desviación típica se han calculado ya en los ejercicios 4.1.*a* y 4.2.*c* del capítulo anterior).

b. No. Eso sólo ocurrirá si la distribución es normal. Podemos comprobarlo calculando el percentil 33: $i = k(n+1)/100 = 33\,(20+1)/100 = 6{,}93$. Hay que trabajar con las posiciones 6ª y 7ª, es decir, con las puntuaciones 14 y 15: $P_{33} = (1-0{,}93)\,14 + (0{,}93)\,15 = 14{,}93$ (distinta de 13).

c. Simplemente multiplicarlas por 5.

d. Llamando X a las puntuaciones transformadas: $X = 5 + 2\,Z_Y$.

e. No. Las puntuaciones típicas no se alteran cuando se efectúan transformaciones lineales. Por tanto, $Z_X = Z_Y$.

f No. Como resultado de esa transformación se obtienen puntuaciones negativas y con decimales. Las puntuaciones negativas no se entienden bien como puntuaciones de una escala a no ser que el cero se tome como referente del centro de la distribución. Y los decimales pueden dar la falsa impresión de que se está midiendo con un elevado nivel de precisión (cosa que en realidad no se está haciendo).

5.2. *a*. Si se suman las puntuaciones directas no se consigue el objetivo porque de ese modo tendría más peso la prueba con mayor media (y si las varianzas fueran muy distintas, no habría manera de saber lo que significa esa puntuación final). Antes de sumar las dos puntuaciones hay que tipificarlas. Llamando Y_F a las puntuaciones finales, Y_1 a las puntuaciones de la primera prueba e Y_2 a las de la segunda, la forma de obtener puntuaciones finales en las que ambas pruebas tengan el mismo peso es la siguiente: $Y_F = Z_{Y_1} + Z_{Y_2}$.

b. $Y_F = \frac{2}{3} Z_{Y_1} + \frac{1}{3} Z_{Y_2}$.

c. Primer candidato: $Z_{(14)} = (14-15)/4{,}51 = -0{,}22$; $Z_{(9)} = (9-8)/3{,}24 = 0{,}31$.

$Y_{\text{final}} = -0{,}22 + 0{,}31 = 0{,}09$.

Segundo candidato: $Z_{(17)} = (17-15)/4{,}51 = 0{,}44$; $Z_{(6)} = (6-8)/3{,}24 = -0{,}62$.

$Y_{\text{final}} = 0{,}44 - 0{,}62 = -0{,}18$.

Por tanto, si ambas pruebas puntúan igual (tienen el mismo peso), el primer sujeto tiene una puntuación final más alta (0,09) que el segundo (−0,18).

d. Primer candidato: $Y_{\text{final}} = (2/3)(-0{,}22) + (1/3)(0{,}31) = -0{,}44$.

Segundo candidato: $Y_{\text{final}} = (2/3)(0{,}44) + (1/3)(-0{,}62) = -0{,}44$.

Por tanto, si la primera prueba puntúa el doble que la segunda (es decir, tiene el doble de peso que la segunda en la puntuación final), el segundo sujeto tiene una puntuación final más alta (0,09) que el primero (−0,04).

e. Llamando Y_F a las puntuaciones finales calculadas en los apartados *a* o *b* y X a las puntuaciones transformadas: $X = 100 + 20\,Z_{Y_F}$.

5.3. *a*. No necesariamente. El valor de una puntuación típica depende de la media y de la desviación típica de la variable. De hecho, el décimo sujeto, que puntúa más alto en el momento *antes* ($Y_{\text{antes}} = 19$) que en el momento *después* ($Y_{\text{después}} = 12$), tiene una puntuación típica más baja en el momento *antes* ($Z_{(\text{antes}=19)} = (19-16)/8{,}5 = 0{,}35$) que en el momento *después* ($Z_{(\text{antes}=12)} = (12-8)/4{,}37 = 0{,}92$). Y lo mismo ocurre con los sujetos 3, 5 y 6 (ver, más abajo, la tabla incluida en el apartado *e*).

b. Sí. En una misma variable, cuanto mayor es una puntuación típica, mayor es también el porcentaje de sujetos que acumula.

c. No necesariamente. Los percentiles se basan únicamente en el orden (mayor, menor) de las puntuaciones. Las puntuaciones típicas tienen en cuenta, no sólo el orden, sino la distancia entre las puntuaciones.

d. Sí. En una misma variable (D_Y), las puntuaciones directas y las puntuaciones típicas expresan cosas equivalentes: cuanto mayor es una puntuación directa, mayor es también su puntuación típica (esto puede comprobarse más abajo, en la tabla incluida en el siguiente apartado).

e. No. La correspondencia que se da entre las puntuaciones típicas de dos variables distintas no es equivalente a la que se da entre sus correspondientes puntuaciones directas. Esto es debido a que cada variable tiene su propia media y su propia desviación típica (ver la respuesta del apartado *a* de este mismo ejercicio). En la siguiente tabla puede apreciarse que no existe correspondencia entre las variables D_Y y D_Z.

Sujetos	Y_{antes}	$Y_{\text{después}}$	D_Y	Z_{D_Y}	Z_{antes}	$Z_{\text{después}}$	D_Z
1	9	10	−1	−1,56	−0,82	0,46	−1,28
2	8	2	6	−0,35	−0,94	−1,37	0,43
3	21	11	10	0,35	0,59	0,69	−0,10
4	14	6	8	0,00	−0,24	−0,46	0,22
5	3	3	0	−1,39	−1,53	−1,14	−0,39
6	17	10	7	−0,17	0,12	0,46	−0,34
7	14	3	11	0,52	−0,24	−1,14	0,91
8	33	15	18	1,73	2,00	1,60	0,40
9	22	8	14	1,04	0,71	0,00	0,71
10	19	12	7	−0,17	0,35	0,91	−0,56

f. No. Una vez tipificadas las puntuaciones directas, las medias valen cero y la diferencia entre ellas también. Y eso no refleja el cambio experimentado entre ambos momentos.

5.4. Sujeto 1: a partir de su puntuación directa $Y = 11$ se obtienen las demás.
Sujeto 2: si $y = 0$, entonces $Y = y + \overline{Y} = 0 + 15 = 15$.
Sujeto 3: si $Z_Y = 0{,}5$, entonces $y = 4(0{,}5) = 2$; $Y = y + \overline{Y} = 2 + 15 = 17$.
Sujeto 4: si $X = 45$, entonces $Z_Y = (45-50)/10 = -0{,}5$; $y = 4(-0{,}5) = -2$; $Y = y + \overline{Y} = -2 + 15 = 13$.
Sujeto 5: su puntuación diferencial vale 4 (pues la suma de las puntuaciones diferenciales vale cero). A partir de ahí se puede proceder como en el sujeto 2: si $y = 4$, entonces $Y = y + \overline{Y} = 4 + 15 = 19$.

Sujetos	Y	y	Z_Y	X
1	11	−4,0	−1,0	40
2	15	0	0	50
3	17	2	0,5	55
4	13	−2,0	−0,5	45
5	19	4	1	60

5.5. *a*. 0,1977; 0,7088; 0,8925 − 0,3409 = 0,5516; 0,2389 + 0,1788 = 0,4177.
b. 0; −0,62; 1,11; 0,28 y 1,26.

5.6.
a. 0,5000.
b. 0,3085.
c. 0,0344.
d. 0,6915.
e. 0,9656.
f. 0,5000.
g. 0,3085.
h. 0,1056.
i. 0,6915.
j. 0,8944.
k. 0,1915.
l. 0,4842.
m. 0,2927.
n. 0,0954.
ñ. 0,1915.
o. 0,4452.
p. 0,2537.
q. 0,0390.
r. 0,3830.
s. 0,6571.
t. 0,6571.
u. 0,9108.
v. 0,5344.
w. 0,1400.
x. 0,8085.
y. 0,8447.
z. 1,0000.

5.7.
a. −3,00.
b. −2,56.
c. −2,33.
d. −1,96.
e. −1,64.
f. −1,28.
g. −0,67.
h. 0,00.
i. 0,67.
j. 1,28.
k. 1,64.
l. 1,96.
m. 2,33.
n. 2,56.
ñ. 3,00.

5.8. No. Aunque la media de esas 5 puntuaciones vale 0, la desviación típica no vale 1, sino 1,58.

5.9. *a*. $Z_{(130)} = (130 - 100)/15 = 2$.

b. $P(\text{CI} > 130) = P(Z > 2) = 0{,}0228$. El 2,28% de los casos.

c. En puntuaciones típicas, entre $Z_{0,025} = -1{,}96$ y $Z_{0,975} = 1{,}96$.

$-1{,}96 = (\text{CI}_1 - 100)/15 \rightarrow \text{CI}_1 = 100 - 1{,}96\,(15) = 70{,}06$.

$1{,}96 = (\text{CI}_2 - 100)/15 \rightarrow \text{CI}_2 = 100 + 1{,}96\,(15) = 129{,}4$.

En puntuaciones directas (CI), entre 70,6 y 129,4.

d. Llamando X a la puntuación transformada: $X = 5 + 1{,}5\,(2) = 8$.

e. Los puntos de corte deben ser los percentiles 5, 25, 75 y 95, los cuales, tratándose de una distribución normal, son, en puntuaciones típicas: $Z_{0,05} = -1{,}65$, $Z_{0,25} = -0{,}67$, $Z_{0,75} = 0{,}67$, $Z_{0,95} = 1{,}65$. Y, en puntuaciones directas (CI):

Percentil 5: $-1{,}65 = (P_5 - 100)/15 \rightarrow P_5 = -1{,}65\,(15) + 100 = 75{,}25$.
Percentil 25: $-0{,}67 = (P_{25} - 100)/15 \rightarrow P_{25} = -0{,}67\,(15) + 100 = 89{,}95$.
Percentil 75: $0{,}67 = (P_{75} - 100)/15 \rightarrow P_{75} = 0{,}67\,(15) + 100 = 110{,}05$.
Percentil 95: $1{,}65 = (P_{95} - 100)/15 \rightarrow P_{95} = 1{,}65\,(15) + 100 = 124{,}75$.

5.10. *a*. Dos horas es justamente el tiempo medio de la distribución. Y en una distribución normal, la media deja por debajo de sí (a la izquierda) el 50% de las puntuaciones.

b. $P(Y < 100) = P[Z < (100 - 120)/20] = P(Z < -1) = 0{,}1587$. Aproximadamente el 16% de los estudiantes (el 15,87%).

c. Hay que calcular el percentil 85. Sabemos (tabla de la curva normal) que $Z_{0,85} = 1{,}04$. Por tanto, $1{,}04 = (Y - 120)/20 \rightarrow Y = 120 + 1{,}04\,(20) = 140{,}8$ minutos.

5.11. $P(Z < -2) + P(Z > 2) = 2F(Z = -2) = 2\,(0{,}0228) = 0{,}0456$. En porcentaje: 4,56%.

5.12. $Z_{0,063} = -1{,}53$; $Z_{0,992} = 2{,}41$.

$-1{,}53 = (147 - \bar{Y})\,/\,S_Y \rightarrow \bar{Y} = 1{,}53\,S_Y + 147$.

$2{,}41 = (541 - \bar{Y})\,/\,S_Y \rightarrow 2{,}41 = [541 - (1{,}53\,S_Y + 147)]/S_Y$.

$2{,}41\,S_Y + 1{,}53\,S_Y = 541 - 147 \rightarrow 3{,}94\,S_Y = 394$.

$S_Y = 394/3{,}94 = 100$.

$\bar{Y} = 1{,}53\,(100) + 147 = 300$.

5.13. *a*. $P(P_{25} < Y_1 < P_{75}) \times P(P_{25} < Y_2 < P_{75}) = P(-0{,}67 < Z < 0{,}67)^2 = (0{,}50)^2 = 0{,}25$.

b. $P(Z > 0) \times P(Z > 0) = (0{,}50)^2 = 0{,}25$.

c. $P(Z > 2) \times P(Z > 2) = F(Z = -2)^2 = 0{,}0228^2 = 0{,}00052$.

5.14. Si la distancia entre $Z = -1$ y $Z = 1$ vale 20 puntos, la distancia entre $Z = 0$ y $Z = 1$ vale 10 puntos. Por tanto, la desviación típica vale 10 puntos (debe tenerse en cuenta que una puntuación típica no es más que el número de desviaciones típicas que una puntuación directa se aleja de su media). Por otro lado, si el percentil 90 ($Z_{0,90} = 1{,}28$) vale 40, entonces

$1{,}28 = (40 - \bar{Y})/10 \rightarrow \bar{Y} = 40 - 10\,(1{,}28) = 27{,}2$.

5.15. Tenemos una variable dicotómica (1 = «en contra», 0 = «a favor») de la que se han realizado $n = 50$ ensayos. Siendo n_1 = «número de personas que se manifiestan en contra de la eutanasia» y π_1 = «pro-

porción de personas en contra de la eutanasia», la variable n_1 sigue el modelo de probabilidad binomial. Como $n > 20$ (límite de la tabla de la distribución binomial), podemos utilizar la aproximación de la binomial a la normal:

$$P(n_1 > 25) = P[Z > (26 - 50 \times 0{,}40)/\sqrt{50(0{,}40)(0{,}60)}] = P(Z > 1{,}73) = 1 - 0{,}9582 = 0{,}042.$$

5.16. *a*. Se han realizado $n = 30$ ensayos de una variable dicotómica (1 = «acierto», 0 = «error»). La variable n_1 es la suma de 30 ensayos de Bernoulli independientes entre sí; por tanto, su distribución de probabilidad es la binomial, con

$$E(n_1) = n\pi_1 = 30(0{,}50) = 15 \quad \text{(ecuación [3.7])}.$$

$$V(n_1) = n\pi_1(1-\pi_1) = 30(0{,}50)(0{,}50) = 7{,}5 \quad \text{(ecuación [3.7])}.$$

b. Puesto que el tamaño muestral es mayor que 20 (límite de la Tabla B), recurrimos a la aproximación de la distribución binomial a la normal:

$$P(n_1 > 20) = P[Z > (21 - 15)/\sqrt{7{,}5}\,] = P(Z > 2{,}19) = 1 - 0{,}9857 = 0{,}0143.$$

5.17. Vamos a calcular la probabilidad de obtener con el tratamiento convencional el resultado obtenido con el experimental. Si esa probabilidad es muy baja, significará que el resultado obtenido con el tratamiento experimental es muy difícil de obtener con el convencional.

Tenemos $n = 40$ ensayos de una variable dicotómica (1 = «recuperación», 0 = «no recuperación») con probabilidad de recuperación $\pi_1 = 0{,}40$ en cada ensayo. Tenemos, por tanto, una variable (n_1 = «número de recuperaciones») que se distribuye binomialmente. Pero, como $n > 20$, vamos a recurrir a la aproximación de la distribución binomial a la normal:

$$P(n_1 > 24) = P[Z > (25 - 40 \times 0{,}40)/\sqrt{40(0{,}40)(0{,}60)}] = P(Z > 1{,}90) = 1 - 0{,}9981 = 0{,}0019.$$

Es muy poco probable (0,0019) que el resultado obtenido con el tratamiento experimental pueda obtenerse con el convencional. Por tanto, parece que el método experimental funciona mejor que el convencional.

5.18. Tenemos una variable dicotómica (1 = «acierto», 0 = «error») de la que se han realizado $n = 30$ ensayos, con n_1 = «número de aciertos». Se trata de calcular el valor del percentil 95 cuando los sujetos responden al azar, es decir, cuando la probabilidad de acierto en cada ensayo vale $\pi_1 = 1/5 = 0{,}20$:

$$Z_{0,95} = 1{,}645 = (n_1 - 30 \times 0{,}20)/\sqrt{30(0{,}20)(1 - 0{,}20)} = (n_1 - 6)/2{,}19.$$

$$n_1 = 6 + 1{,}645(2{,}19) = 9{,}6.$$

El criterio establecido lleva a colocar el punto de corte para el aprobado en 9,6 aciertos. Es decir, será necesario acertar al menos 10 preguntas para aprobar.

6

Las distribuciones muestrales

Aunque ya hemos hablado del concepto de *distribución de probabilidad*, conviene volver sobre él de nuevo. Su rol dentro de la estadística es lo bastante importante como para que todo analista de datos esté especialmente familiarizado con él.

Recordemos que una tabla de frecuencias contiene información sobre la forma de la distribución de una variable categórica (ver Capítulo 3). Más concretamente, las frecuencias relativas (proporciones) de una tabla de frecuencias indican con qué probabilidad cabe esperar encontrar cada uno de los valores de la variable. Si sabemos, por ejemplo, que el 30% de las personas de un determinado colectivo son fumadoras, sabemos que la probabilidad de que una persona seleccionada al azar de ese colectivo sea fumadora vale 0,30. En este sentido, una tabla de frecuencias relativas puede concebirse como una distribución de probabilidad, es decir, como una distribución que contiene todos los valores de la variable junto con la probabilidad asociada a cada uno de ellos.

Esto mismo vale también para las variables cuantitativas. Por ejemplo, cuando se mide la altura en un grupo de sujetos y con los valores obtenidos se construye un histograma (ver Capítulo 4), estamos captando la forma de la distribución; y las frecuencias relativas asociadas a cada valor o rango de valores están indicando con qué probabilidad cabe esperar encontrar cada uno de ellos.

Ahora bien, en ambos casos estamos hablando de **distribuciones empíricas**, es decir, de distribuciones construidas a partir de los datos observados. Pero sabemos que también existen **distribuciones teóricas** como la binomial o la normal. Son distribuciones que, aunque no están generadas a partir de los datos sino a partir de una función matemática, son representaciones de los datos que tienen la enorme utilidad de ayudar a interpretarlos mejor.

Qué es una distribución muestral

Las distribuciones de probabilidad, las empíricas y las teóricas, son importantes porque contienen toda la información relativa a una variable. Una distribución de probabilidad, es decir, los valores de la variable combinados con la frecuencia con la que se da cada uno de ellos, informa sobre el centro, la dispersión y la forma de la distribución; y eso, ya lo sabemos, es todo lo que necesitamos para caracterizar una variable.

Lo que interesa destacar en este momento es que *los estadísticos son variables*. Y como variables que son, tienen, al igual que cualquier otra variable, su propia distribución de probabilidad. Pues bien, el término **distribución muestral** se refiere a la distribución de probabilidad (o de densidad) de un estadístico. Por tanto, una distribución muestral puede definirse como:

Una distribución teórica que asigna una probabilidad (o una densidad) concreta a cada uno de los valores que puede tomar un estadístico en todas las muestras del mismo tamaño que es posible extraer de una determinada población.

El concepto de distribución desempeña un rol esencial en el análisis de datos. Como tal, merece una atención algo detallada.

Un caso concreto

Uno de los estadísticos más útiles y utilizados en el contexto de la inferencia estadística es la media aritmética: $\bar{Y}$. Ya sabemos que, en cuanto estadístico que es, su valor concreto depende de la muestra concreta en la que se calcula.

De una población cualquiera es posible extraer más de una muestra de tamaño n (en una población infinita es posible extraer infinitas muestras de cualquier tamaño). Si en cada una de esas muestras calculamos $\bar{Y}$, podremos comprobar que no siempre toma el mismo valor, sino que varía de una muestra a otra. Estos posibles valores del estadístico $\bar{Y}$ constituyen su distribución muestral.

Veamos con un sencillo ejemplo cómo se construye la distribución muestral del estadístico $\bar{Y}$. Imaginemos una población formada por $N = 5$ puntuaciones: $Y_i = \{1, 2, 3, 4, 5\}$. Si de esa población se seleccionan aleatoriamente y con reposición todas las posibles muestras de tamaño $n = 2$, se obtendrán $N^n = 5^2 = 25$ muestras distintas (variaciones con repetición de N elementos tomados de n en n), todas las cuales tendrán la misma probabilidad de ser elegidas: 1/25 (principio de indiferencia; ver, en el Capítulo 2, el apartado *Probabilidad*). Si ahora se calcula en cada una de esas 25 muestras el estadístico $\bar{Y}$, se obtendrán los resultados que muestra la Tabla 6.1.

La Tabla ofrece las 25 muestras posibles y el valor que toma el estadístico $\bar{Y}$ en cada una de ellas[1]. En estos resultados pueden apreciarse diferentes cosas. Por ejemplo, aunque sólo en una de las 25 muestras se obtiene el valor $\bar{Y} = 1$, el valor $\bar{Y} = 2{,}5$ se obtiene en cuatro de ellas. Esto significa que el estadístico $\bar{Y}$ puede tomar el mismo valor en más de una muestra, pero también significa (y esto es lo realmente interesante) que, aunque las 25 muestras son equiprobables (todas tienen la misma probabilidad de ser elegidas), los posibles valores del estadístico $\bar{Y}$ no lo son: hay unos valores de $\bar{Y}$ que son más probables que otros porque unos pueden obtenerse en un mayor número de muestras que otros. Puede comprobarse en la tabla que, efectivamente, hay más muestras en las que se obtiene, por ejemplo, $\bar{Y} = 2{,}5$ que $\bar{Y} = 1{,}5$. Estas consideraciones sugieren que los datos de la Tabla 6.1 pueden resumirse tal como se muestra en la Tabla 6.2. En ella aparecen los diferentes valores que puede tomar el estadístico $\bar{Y}$ junto con la probabilidad (frecuencia relativa) asociada a cada uno de ellos, es decir, la *dis-*

[1] El ejemplo utilizado en este apartado es a todas luces un ejemplo irreal sin ningún tipo de relación con la investigación empírica. Sin embargo, su simplicidad le confiere la virtud de permitir explicar fácilmente el importantísimo concepto de distribución muestral.

tribución muestral de la media (en concreto, la distribución muestral correspondiente a muestras de tamaño $n = 2$ extraídas de una población de $N = 5$ elementos).

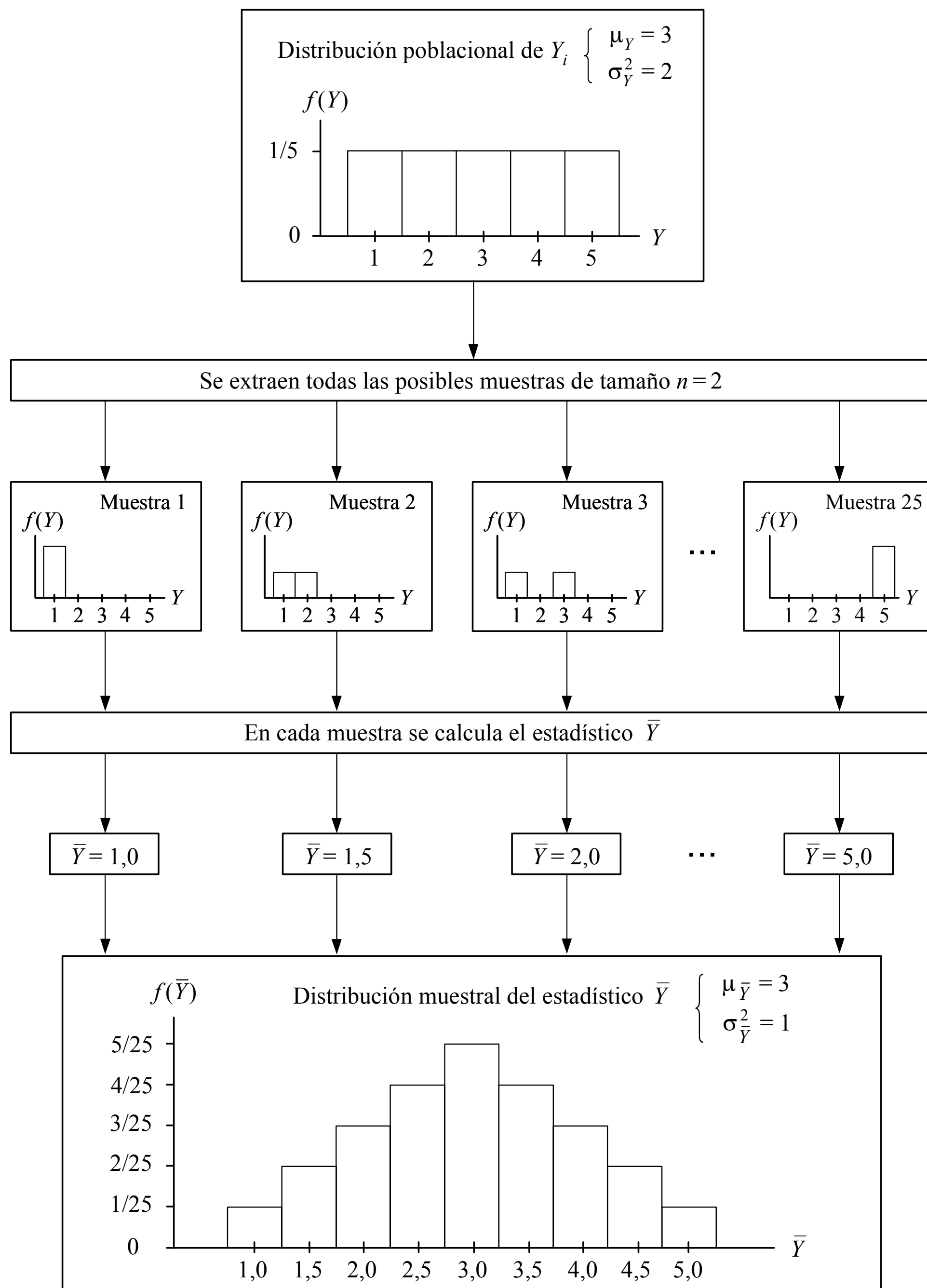
Tabla 6.1. Muestras de tamaño n = 2 que es posible extraer con reposición de una población de N = 5 elementos, valor del estadístico media ($\bar{Y}$) en cada una de ellas y probabilidad asociada (f) a cada valor $\bar{Y}$

Muestras	*Valores muestrales*	$\bar{Y}$	$f(\bar{Y})$
1	1, 1	1,0	1/ 25
2	1, 2	1,5	1/ 25
3	1, 3	2,0	1/ 25
4	1, 4	2,5	1/ 25
5	1, 5	3,0	1/ 25
6	2, 1	1,5	1/ 25
7	2, 2	2,0	1/ 25
8	2, 3	2,5	1/ 25
9	2, 4	3,0	1/ 25
10	2, 5	3,5	1/ 25
11	3, 1	2,0	1/ 25
12	3, 2	2,5	1/ 25
13	3, 3	3,0	1/ 25
14	3, 4	3,5	1/ 25
15	3, 5	4,0	1/ 25
16	4, 1	2,5	1/ 25
17	4, 2	3,0	1/ 25
18	4, 3	3,5	1/ 25
19	4, 4	4,0	1/ 25
20	4, 5	4,5	1/ 25
21	5, 1	3,0	1/ 25
22	5, 2	3,5	1/ 25
23	5, 3	4,0	1/ 25
24	5, 4	4,5	1/ 25
25	5, 5	5,0	1/ 25

Tabla 6.2. Distribución muestral de la media ($\bar{Y}$) formada a partir de los datos de la Tabla 6.1

N º de muestras	$\bar{Y}$	$f(\bar{Y})$
1	1,0	1/ 25
2	1,5	2/ 25
3	2,0	3/ 25
4	2,5	4/ 25
5	3,0	5/ 25
4	3,5	4/ 25
3	4,0	3/ 25
2	4,5	2/ 25
1	5,0	1/ 25
		1

Figura 6.1. Procedimiento seguido para obtener la distribución muestral del estadístico *media* ($\bar{Y}$) a partir de todas las muestras de tamaño n = 2 que es posible extraer, con reposición, de una población formada por los números 1, 2, 3, 4 y 5 (figura adaptada de Kirk, 1978, pág. 205)



La distribución muestral de la media puede obtenerse, como veremos a continuación, por procedimientos puramente matemáticos, sin necesidad de tener que extraer todas las posibles muestras de tamaño n (cosa que, por otra parte, es impensable con una población infinita). Sin embargo, obtener una distribución muestral a partir de todos sus posibles valores en muestras de tamaño n tiene la ventaja de ayudar a reparar en ciertos detalles que de otro modo podrían pasar desapercibidos. En la Tabla 6.2 puede comprobarse, por ejemplo, que el valor de $\bar{Y}$ que tiene asociada la probabilidad más alta es 3, valor que, curiosamente, coincide con la media de la población:

$$\mu_Y = (1+2+3+4+5)/5 = 3$$

También puede comprobarse, por ejemplo, que la distribución tiene forma simétrica (con el centro en el valor 3); y que los valores más alejados del centro son los que tienen asociadas probabilidades más pequeñas.

Continuando con nuestra aproximación intuitiva al concepto de distribución muestral hemos elaborado el gráfico que muestra la Figura 6.1. Este gráfico ilustra cómo se llega a la distribución muestral del estadístico $\bar{Y}$ seleccionando todas las posibles muestras de tamaño 2 de una población de 5 elementos equiprobables y calculando el estadístico $\bar{Y}$ en cada una de ellas. El gráfico también ilustra un hecho peculiar: la forma de la distribución muestral de $\bar{Y}$ no se parece a la de la distribución poblacional de Y: mientras que ésta es uniforme (los 5 elementos poblacionales son equiprobables), la distribución muestral de $\bar{Y}$ sigue siendo simétrica pero con forma de campana (se parece a una distribución normal).

Otro caso concreto

Lo que acabamos de hacer con el estadístico $\bar{Y}$ puede hacerse con cualquier otro estadístico imaginable: cualquier otro estadístico que decidamos calcular tomará diferentes valores en las diferentes muestras de tamaño n que puedan extraerse de una población y eso permitirá determinar su distribución de probabilidad, es decir, su distribución muestral.

Consideremos ahora una población de $N = 10$ personas formada por 6 hombres y 4 mujeres, y extraigamos de ella aleatoriamente y con reposición muestras de tamaño $n = 3$. Como el muestreo realizado es con reposición, el primer elemento muestral puede ser cualquiera de los 10 posibles; el segundo, también cualquiera de los 10 posibles; y el tercero, también cualquiera de los 10 posibles. Por tanto, aplicando el principio fundamental de la combinatoria (ver Apéndice 1), el número de muestras posibles será de $10 \times 10 \times 10 = 1.000$ (es decir, variaciones con repetición de 10 elementos tomados de 3 en 3: $N^n = 10^3 = 1.000$). Estas 1.000 muestras posibles son las que recoge la Tabla 6.3.

En este nuevo escenario ya no es posible definir el estadístico media ($\bar{Y}$) pues no tiene sentido calcular la media de HHH, o de HHM. Pero sí se puede definir, por ejemplo, el estadístico n_H = «número de hombres en la muestra». Dependiendo de la muestra elegida, n_H podrá tomar los valores 0, 1, 2 o 3; es decir, en una muestra cualquiera de las 1.000 posibles puede ocurrir que no haya ningún hombre, que haya uno, que haya dos o que los tres sean hombres. La Tabla 6.3 recoge estos posibles valores de n_H.

También es posible definir otro estadístico estrechamente relacionado con n_H, en concreto: P_H = «proporción de hombres en la muestra». Los valores de P_H dependen de los que tome n_H, pues P_H no es más que una transformación de n_H. La Tabla 6.3 también muestra los posibles valores de P_H.

Tabla 6.3. Muestras de tamaño n = 3 que es posible extraer con reposición de una población de N = 10 personas de las que 6 son hombres y 4 son mujeres (H = hombre, M = mujer)

1ª extracción	2ª extracción	3ª extracción	N º de muestras[2]	N º de hombres (n_H)	Proporción de hombres (P_H)
H	H	H	216	3	3/3
H	H	M	144	2	2/3
H	M	H	144	2	2/3
M	H	H	144	2	2/3
H	M	M	96	1	1/3
M	H	M	96	1	1/3
M	M	H	96	1	1/3
M	M	M	64	0	1/3
			1.000		

Hemos definido dos estadísticos (n_H y P_H), pero todavía no hemos definido ninguna distribución muestral. Para ello es necesario asociar a cada valor de n_H o de P_H su correspondiente probabilidad (frecuencia relativa). Y la Tabla 6.3 contiene toda la información necesaria para hacerlo. Con esa información se han construido las distribuciones muestrales que recoge la Tabla 6.4[3].

En el apartado anterior hemos visto que el valor más probable del estadístico media ($\overline{Y}$), es decir, su valor esperado, coincidía con su correspondiente parámetro (μ_Y). Con el estadístico proporción ocurre lo mismo (en realidad, una proporción no es otra cosa que una media). Recordemos que la población está formada por 10 personas de las que 6 son hombres. Es decir, la proporción de hombres en la población es $\pi_H = 6/10 = 0{,}6$. Pues bien, en la distribución muestral de P_H (ver Tabla 6.4) ocurre que el valor esperado del estadístico P_H es justamente el parámetro π_H:

$$E(P_H) \;=\; \sum P_H\, f(P_H) \;=\; (0/3)\,0{,}064 + (1/3)\,0{,}288 + (2/3)\,0{,}432 + (3/3)\,0{,}216 \;=\; 060$$

De nuevo puede constatarse que una distribución muestral contiene la probabilidad asociada a cada uno de los valores que puede tomar un estadístico en todas las posibles muestras de tamaño n. Así, por ejemplo, si de una población formada por 10 personas, 6 hombres y 4 mujeres, extraemos aleatoriamente una muestra de tamaño 3, sabemos que lo más probable (0,432)

[2] Los resultados de la columna "número de muestras" pueden obtenerse fácilmente aplicando el teorema fundamental de la combinatoria (ver Apéndice 2). Así, por ejemplo, en la muestra formada por un hombre, una mujer y otro hombre (HMH), tendremos que el primer suceso (H) puede ocurrir de 6 maneras (pues hay 6 hombres en la población); el segundo suceso (M) puede ocurrir de 4 maneras (pues hay 4 mujeres en la población); y el tercer suceso (H) puede ocurrir de 6 maneras (pues hay 6 hombres en la población y el muestreo es con reposición). Por tanto, los tres sucesos juntos pueden ocurrir de $6 \times 4 \times 6 = 144$ maneras.

[3] La columna $f(n_H) = f(P_H)$ ofrece las probabilidades (frecuencias relativas) asociadas a los estadísticos n_H y P_H, respectivamente. Así, por ejemplo, en la Tabla 6.3 hay 3(144) = 432 muestras en las que el número de hombres es 2. Esto está reflejado en la Tabla 6.4 en el hecho de que la probabilidad de que n_H tome el valor 2 o P_H el valor 2/3 vale 432/1.000 = 0,432. Obviamente, $f(n_H)$ es igual a $f(P_H)$, pues hablar de la probabilidad de que en una muestra de 3 personas aparezcan $n_H = 2$ hombres es exactamente lo mismo que hablar de la probabilidad de que, en esa misma muestra, la proporción de hombres sea $P_H = 2/3$.

es que la proporción de hombres en esa muestra valga $P_H = 2/3 = 0{,}67$ (o, lo que es lo mismo, $n_H = 2$). También sabemos, por ejemplo, que es muy poco probable (0,064) que ninguna de las tres personas elegidas sea hombre ($n_H = P_H = 0$). Etcétera.

Tabla 6.4. Distribución muestral de los estadísticos n_H = «número de hombres» y P_H = «proporción de hombres» obtenida a partir de los datos de la Tabla 6.3

n_H	P_H	$f(n_H) = f(P_H)$
0	0/3	64 / 1.000 = 0,064
1	1/3	228 / 1.000 = 0,228
2	2/3	432 / 1.000 = 0,432
3	3/3	216 / 1.000 = 0,216
		1

El caso general

Con poblaciones y muestras pequeñas como las de los dos ejemplos propuestos resulta relativamente sencillo conocer la distribución muestral de un estadístico. Pero las poblaciones con las que habitualmente se trabaja no son, ni mucho menos, tan pequeñas como las de estos ejemplos. Más bien al contrario, las poblaciones que interesa estudiar suelen ser muy grandes e incluso, en ocasiones, infinitas. Y esto significa que, para obtener la distribución muestral de un estadístico, no siempre resulta posible extraer todas las posibles muestras de tamaño n.

Sin embargo, el concepto de distribución muestral sigue siendo el mismo cualesquiera que sean el tamaño de la población y de la muestra. En una población infinita, la distribución muestral de, por ejemplo, el estadístico $\overline{Y}$ sigue siendo la distribución resultante de extraer infinitas muestras de tamaño n y calcular en todas ellas $\overline{Y}$. Por supuesto, no es posible extraer las infinitas muestras de tamaño n de una población infinita, pero eso no significa que haya que renunciar a conocer la distribución muestral de un estadístico. Existen procedimientos matemáticos capaces de captar con precisión las características de una distribución muestral.

Sabemos que una distribución muestral es la distribución de probabilidad de un estadístico. También sabemos que un estadístico es una variable aleatoria. Por tanto, la distribución muestral de un estadístico puede quedar caracterizada del mismo modo que la distribución de cualquier variable aleatoria, a saber, haciendo explícitos su *centro*, su *dispersión* y su *forma*.

En el contexto de las distribuciones muestrales, el **centro** habitualmente utilizado es la media o *valor esperado*; para cuantificar la **dispersión** suele utilizarse la varianza (o su raíz cuadrada, la desviación típica, la cual cambia de nombre cuando se refiere a un estadístico pasando a llamarse *error típico*; por ejemplo, a la desviación típica del estadístico *media* se le llama *error típico de la media*); y para determinar la **forma** de la distribución se recurre a distribuciones teóricas como la binomial o la normal (ya estudiadas) y a otras derivadas de ellas (que estudiaremos en su momento).

Aunque en los próximos capítulos iremos conociendo algunas de las distribuciones muestrales más utilizadas, en éste vamos a centrarnos en dos de ellas para terminar de afianzar el concepto. En concreto, vamos a estudiar las distribuciones muestrales del estadístico *media* y del estadístico *proporción*, justamente los dos estadísticos utilizados en los dos apartados anteriores para ofrecer una aproximación intuitiva al concepto de distribución muestral.

Distribución muestral del estadístico *media*

Acabamos de ver cómo llegar a conocer la distribución muestral de la media a partir de todas las muestras de tamaño n que es posible extraer de una determinada población. Sin embargo, la distribución muestral de la media también puede conocerse sin necesidad de extraer una sola muestra. Sabemos que la media aritmética de n observaciones independientes es, por definición,

$$\bar{Y} = \frac{\sum Y_i}{n} = \frac{1}{n}Y_1 + \frac{1}{n}Y_2 + \cdots + \frac{1}{n}Y_n$$

A partir de esta fórmula es fácil deducir (ver Apéndice 6, al final de este capítulo) el valor esperado y la varianza del estadístico $\bar{Y}$:

$$E(\bar{Y}) = \mu_Y \qquad [6.1]$$
$$V(\bar{Y}) = \sigma^2_{\bar{Y}} = \sigma^2_Y/n \qquad \rightarrow \qquad \sigma_{\bar{Y}} = \sigma_Y/\sqrt{n}$$

Por tanto, el ***centro*** de la distribución muestral de $\bar{Y}$ es el mismo que el de la distribución poblacional de Y. Y la ***varianza*** de la distribución muestral de $\bar{Y}$ es la varianza poblacional de Y dividida por el tamaño de la muestra[4]. Además, si las variables Y_1, Y_2, ..., Y_n de las que se obtiene $\bar{Y}$ se distribuyen normalmente con parámetros μ_Y y σ_Y, entonces la ***forma*** de la distribución muestral de $\bar{Y}$ es también normal con parámetros μ_Y y $\sigma_{\bar{Y}}$, lo cual se representa mediante

$$\bar{Y} \sim N(\mu_Y, \sigma_{\bar{Y}}) \qquad [6.2]$$

Y todavía más. De acuerdo con el *teorema del límite central* (ver, en el Capítulo 5, el apartado *Curva normal*), si Y_1, Y_2, ..., Y_n son variables aleatorias independientes e igualmente distribuidas (cualquiera que sea la distribución) con parámetros μ_Y y σ_Y, ambos finitos, la distribución muestral de $\bar{Y}$ tiende a la normalidad, con parámetros μ_Y y $\sigma_{\bar{Y}}$, a medida que n va aumentando[5]. Ahora bien, si $\bar{Y}$ es una variable aleatoria que, bajo las mencionadas circunstancias, se distribuye normalmente, entonces, bajo esas mismas circunstancias, la transformación de la variable $\bar{Y}$ en puntuaciones Z se distribuye normalmente con media igual a cero y desviación típica igual a uno (ver Capítulo 5); es decir:

$$Z = \frac{\bar{Y} - \mu_Y}{\sigma_{\bar{Y}}} \sim N(0, 1) \qquad [6.3]$$

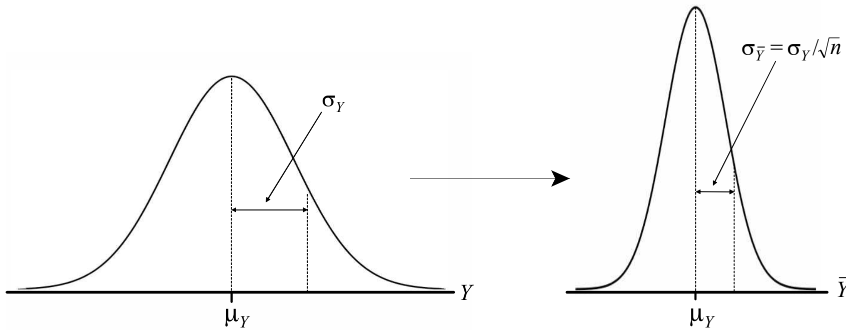
[4] Con muestreo aleatorio *sin reposición* en población finita se puede seguir asumiendo que las variables Y_1, Y_2, ..., Y_n tienen la misma distribución, pero ya no es posible asumir que son **independientes** (como lo serían si el muestreo es *con reposición*). Con poblaciones pequeñas, esto tiene sus consecuencias sobre la varianza (y, por tanto, sobre el error típico) de la distribución muestral de la media, que pasa a ser $\sigma^2_{\bar{Y}} = (\sigma^2_Y/n)[(N-n)/(N-1)]$, donde N se refiere al tamaño de la población y n al de la muestra (ver Amón, 1984, págs. 221-222). Lógicamente, a medida que N vaya aumentando, el término corrector $(N-n)/(N-1)$ irá tendiendo a 1, de manera que, si la población es lo bastante grande, la varianza resultante de muestrear sin reposición una población finita será prácticamente idéntica a la resultante de muestrear con reposición.

[5] La aproximación es muy rápida (tamaños muestrales pequeños) si la distribución original no es muy asimétrica. Cuanto más asimétrica es la distribución original, mayor necesita ser el tamaño muestral para que la distribución muestral de la media se aproxime a la normal.

Esto significa que es posible utilizar la transformación Z y la distribución normal tipificada (Tabla C del apéndice final) para conocer las probabilidades asociadas a cada uno de los diferentes valores del estadístico $\bar{Y}$ en su distribución muestral.

En este momento es importante aclarar una cuestión que suele generar bastante confusión al trabajar por primera vez con la distribución muestral de $\bar{Y}$. Es importante no confundir la distribución original de la variable Y con la distribución muestral del estadístico $\bar{Y}$: si la variable Y se distribuye normalmente con valor esperado μ_Y y desviación típica σ_Y (ver Figura 6.2, curva de la izquierda), entonces el estadístico $\bar{Y}$ se distribuye normalmente con el mismo valor esperado, μ_Y, pero con diferente desviación típica (error típico): $\sigma_{\bar{Y}} = \sigma_Y/\sqrt{n}$ (ver Figura 6.2, curva de la derecha). La Figura 6.2 muestra con claridad que la distribución original de la variable Y (curva de la izquierda) es mucho más ancha (más dispersa) que la distribución muestral de $\bar{Y}$ (curva de la derecha). Esto significa que, para conocer las probabilidades asociadas a los valores de la variable Y hay que tener presente que la desviación típica de su distribución es σ_Y; mientras que para conocer las probabilidades asociadas al estadístico $\bar{Y}$ (que es lo que realmente nos interesa aquí) hay que tener presente que la desviación típica (error típico) de su distribución es $\sigma_{\bar{Y}} = \sigma_Y/\sqrt{n}$.

Figura 6.2. Distribución de la variable Y (izquierda) y distribución muestral del estadístico $\bar{Y}$ (derecha)



Pero esta estrategia tiene un problema: para obtener la transformación propuesta en [6.3] es necesario utilizar $\sigma_{\bar{Y}}$, lo cual exige conocer el valor del parámetro σ_Y (ver [6.1]); y ocurre que, por lo general, el valor de ese parámetro es desconocido. Por supuesto, siempre es posible sustituir σ_Y por su correspondiente valor muestral S_Y (ver, en el siguiente capítulo, el apartado *Estimación puntual*). Con muestras grandes, sustituir σ_Y por S_Y no constituye un problema importante porque el grado de parecido entre ambos valores suele ser tan alto que las cosas no cambian demasiado (es decir, la sustitución de σ_Y por S_Y no suele alterar de forma importante ni el valor de las puntuaciones Z ni a la forma de su distribución). Sin embargo, con muestras pequeñas, esa sustitución tiene consecuencias que no pueden pasarse por alto. De hecho, al sustituir σ_Y por S_Y, la transformación[6]

$$T = \frac{\bar{Y} - \mu_Y}{\hat{\sigma}_{\bar{Y}}} \qquad \left(\text{con } \hat{\sigma}_{\bar{Y}} = \frac{S_Y}{\sqrt{n}}\right) \qquad [6.4]$$

[6] Nótese que esta transformación es idéntica a la transformación Z definida en [6.3] excepto en un detalle: en la transformación Z se utiliza el parámetro σ_Y; en la transformación T se utiliza el estadístico S_Y (ver Apéndice 9).

ya no se distribuye normalmente, sino según el modelo de probabilidad *t* de *Student* con $n - 1$ grados de libertad (ver Apéndice 5), lo cual se representa mediante:

$$T \sim t_{n-1} \quad [6.5]$$

Por tanto, con σ_Y desconocida y muestras pequeñas, las probabilidades asociadas al estadístico $\overline{Y}$ en su distribución muestral pueden conocerse utilizando la transformación *T* y la distribución *t* de *Student*.

La distribución *t* de *Student* (se describe con detalle en el Apéndice 5) es muy parecida a la distribución normal tipificada (es simétrica y su media vale cero) pero acumula más casos en las colas. Alguno de sus valores están tabulados en la Tabla E del Apéndice final. Los valores *t* están en el interior de la tabla; las cabeceras de las filas contienen los grados de libertad (*gl*) de la distribución; las cabeceras de las columnas muestran la proporción de área que acumula cada valor *t* del interior de la tabla. Estas proporciones acumuladas permiten conocer la probabilidad asociada a cada valor *t*. Así, por ejemplo, el cuantil 95 (o valor *t* que acumula una probabilidad de 0,95) con 50 grados de libertad vale 1,676, es decir,

$$t_{50;\,0{,}95} = 1{,}676$$

La Tabla D sólo ofrece los valores *t* ubicados en las colas de la distribución: del percentil 0,1 al 10 y del 90 al 99,9 (los valores centrales no suelen utilizarse para nada).

Cuando la letra *t* se refiere a la distribución de *Student*, únicamente lleva un subíndice que indica los grados de libertad de la distribución; cuando la letra *t* se refiere a un valor concreto de la distribución de *Student* lleva dos subíndices: el primero indica los grados de libertad de la distribución; el segundo, la probabilidad que acumula el valor *t*.

A medida que *n* vaya aumentando, S_Y se irá pareciendo más y más a σ_Y y cada vez con menor variabilidad (es decir, con menor error típico[7]); con *n* tendiendo a infinito: $S_Y \rightarrow \sigma_Y$. Consecuentemente, a medida que *n* vaya aumentando, *T* se irá pareciendo más y más a *Z*. Por tanto, con tamaños muestrales grandes, los resultados obtenidos con *Z* y *T* serán tan parecidos[8] que podrá utilizarse *Z* en lugar de *T* aunque se desconozca el valor de σ_Y.

Ejemplo. Distribución muestral del estadístico media

Trabajar con distribuciones muestrales requiere estar familiarizado con dos tipos de tareas. La primera de ellas consiste en obtener la distribución muestral de un estadístico a partir de las diferentes muestras de tamaño *n* que es posible extraer de una determinada población (lógicamente, esto hay que aprender a hacerlo con poblaciones y muestras muy pequeñas); el objetivo de esta tarea es ayudar a comprender a fondo el concepto de distribución muestral. Los ejercicios 6.1 a 6.4, 6.15, 6.22 y 6.23, al final del capítulo, se centran en esta tarea.

[7] Ver, en el apartado *Estimación puntual* del próximo capítulo, el concepto de *consistencia* como propiedad de un estimador (nota a pie de página número 1). Y en el Apéndice 9, el apartado sobre la *distribución muestral de la varianza*.

[8] Quizá pueda uno formarse una buena idea del grado de parecido entre la distribución normal y la distribución *t* si se tiene en cuenta que la diferencia *más grande* entre las probabilidades acumuladas de ambas distribuciones es de 5 milésimas con 30 grados de liberad, de 3 milésimas con 50 grados de libertad y de menos de dos milésimas con 100 grados de libertad.

La segunda tarea con la que hay que familiarizarse se refiere a los cálculos necesarios para transformar un estadístico en una variable con distribución de probabilidad conocida (por ejemplo, transformar $\overline{Y}$ en Z o en T) y al manejo de las herramientas (tablas y programas informáticos) necesarias para conocer las probabilidades asociadas a los valores transformados. Este ejemplo muestra cómo realizar esta segunda tarea.

Supongamos que la población de estudiantes universitarios se distribuye normalmente con media 100 y desviación típica 15 en una escala de inteligencia espacial (Y). Si extraemos una muestra aleatoria de 50 estudiantes: (1) ¿cuál es la probabilidad de obtener una media de 104 o mayor?, (2) ¿cuál es la probabilidad de obtener una media comprendida entre 96 y 104?

Puesto que la población de partida, es decir, la distribución de la variable Y, es normal y el valor de σ_Y conocido, la transformación Z basada en [6.3] se distribuirá según el modelo de probabilidad normal con media 0 y desviación típica 1, es decir:

$$Z = \frac{\overline{Y} - \mu_Y}{\sigma_{\overline{Y}}} = \frac{\overline{Y} - \mu_Y}{\sigma_Y/\sqrt{n}} = \frac{\overline{Y} - 100}{15/\sqrt{50}} = \frac{\overline{Y} - 100}{2{,}12} \sim N(0,\, 1)$$

En consecuencia,

(1) $P(\overline{Y} \geq 104) = P(Z \geq \dfrac{104 - 100}{2{,}12}) = P(Z \geq 1{,}89) = 1 - F(1{,}89) = 1 - 0{,}9706 = 0{,}0294.$

Para encontrar la probabilidad que queda por encima del valor $Z = 1{,}89$, en la tabla de la curva normal (Tabla C del Apéndice final) se busca la probabilidad acumulada hasta 1,89 y se resta de uno (o, alternativamente, se busca directamente la probabilidad acumulada hasta $-1{,}89$): $1 - F(1{,}89) = F(-1{,}89) = 0{,}0294$.

(2) $P(96 \leq \overline{Y} \leq 104) = P(\dfrac{96 - 100}{2{,}12} \leq Z \leq \dfrac{104 - 100}{2{,}12}) = P(-1{,}89 \leq Z \leq 1{,}89) =$

$$= F(1{,}89) - F(-1{,}89) = 0{,}9706 - 0{,}0294 = 0{,}9412.$$

El resultado del primer apartado indica que es muy poco probable (0,0284) encontrar medias mayores que 104. El resultado del segundo apartado indica que es muy probable (0,9412) encontrar medias comprendidas entre 96 y 104.

Supongamos ahora que la población de estudiantes universitarios se distribuye normalmente con media 100 y desviación típica desconocida en una escala de inteligencia espacial (Y). Supongamos además que en una muestra aleatoria de 20 estudiantes obtenemos una desviación típica insesgada de 19,5. En este escenario, ¿cuál es la probabilidad de que la media de esa muestra sea mayor que 103?

Este ejemplo es parecido al anterior, pero, al no conocer la desviación típica poblacional, no es posible utilizar la transformación Z. No obstante, como la población de partida es normal, sabemos que la transformación T propuesta en [6.4] se distribuye según el modelo de probabilidad t de *Student* con $n - 1 = 20 - 1 = 19$ grados de libertad, es decir,

$$T = \frac{\overline{Y} - \mu_Y}{\hat{\sigma}_{\overline{Y}}} = \frac{\overline{Y} - \mu_Y}{S_Y/\sqrt{n}} = \frac{\overline{Y} - 100}{19{,}5/\sqrt{20}} = \frac{\overline{Y} - 100}{4{,}36} \sim t_{19}$$

En consecuencia,

$$P(\overline{Y} \geq 103) \; = \; P(T \geq \frac{103\,-100}{4{,}36}) \; = \; P(T \geq 0{,}688)$$

Es decir, la probabilidad de encontrar medias mayores que 103 en su distribución muestral es igual a la probabilidad de encontrar valores T mayores que 0,688 en la distribución t de *Student* con 19 grados de libertad. En la Tabla D del Apéndice final puede comprobarse que el valor 0,688 acumula, aproximadamente, una probabilidad de 0,75. Por tanto, en este nuevo escenario, la probabilidad de obtener una media muestral mayor que 103 vale 0,25.

Distribución muestral del estadístico *proporción*

Consideremos una población cualquiera y una variable dicotómica, es decir, una variable categórica que sólo puede tomar dos valores (acierto-error, tratados-no tratados, recuperados-no recuperados, verdadero-falso, etc.). Llamemos a esos dos valores *éxito* y *fracaso*, y π_1 a la proporción de *éxitos* en la población. Si extraemos muestras aleatorias de tamaño n y, en cada muestra, definimos la variable n_1 = «número de éxitos en las n extracciones», tendremos un estadístico distribuido, si π_1 permanece constante en cada extracción, según el modelo de probabilidad *binomial* (ver, en el Capítulo 3, el apartado *Distribución binomial*) con parámetros n y π_1. Esto se representa, recordemos, mediante

$$n_1 \quad \sim \quad B(n,\, \pi_1)$$

Además, dadas las características del estadístico n_1 (suma de n ensayos independientes de Bernoulli), puede demostrarse que su valor esperado y su varianza son los siguientes (ver notas a pie de página números 6 y 7 del Capítulo 3):

$$E(n_1) \; = \; \mu_{n_1} \; = \; n\pi_1 \qquad [6.6]$$

$$V(n_1) \; = \; \sigma^2_{n_1} \; = \; n\,\pi_1(1-\pi_1) \qquad \rightarrow \qquad \sigma_{n_1} \; = \; \sqrt{n\,\pi_1(1-\pi_1)}$$

Es decir, el ***centro*** de la distribución muestral del estadístico n_1 = «número de éxitos en las n extracciones» es n veces la proporción teórica de éxitos. Su ***varianza*** se obtiene multiplicando por n el producto de la proporción teórica de éxitos y su valor complementario. Y su ***forma*** se ajusta a la distribución binomial con parámetros n y π_1.

Si ahora definimos el estadístico $P_1 = n_1/n$ = «proporción de éxitos en las n extracciones» habremos definido un nuevo estadístico (que, en realidad, no es otra cosa que una *media*) que se distribuye exactamente igual que n_1; es decir: $P_1 \sim B(n,\, \pi_1)$. Y como P_1 no es más que una transformación lineal de n_1, su valor esperado y su varianza serán:

$$E(P_1) \; = \; \mu_{P_1} \; = \; \frac{1}{n}\,E(n_1) \; = \; \frac{1}{n}\,n\,\pi_1 \; = \; \pi_1 \qquad [6.7]$$

$$V(P_1) \; = \; \sigma^2_{P_1} \; = \; \frac{1}{n^2}\,\sigma^2_{n_1} \; = \; \frac{n\,\pi_1(1-\pi_1)}{n^2} \; = \; \frac{\pi_1(1-\pi_1)}{n} \qquad \rightarrow \qquad \sigma_{P_1} = \; \sqrt{\pi_1(1-\pi_1)/n}$$

Es decir, el ***centro*** de la distribución muestral del estadístico P_1 = «proporción de éxitos en n extracciones» coincide con la proporción teórica de éxitos. Su ***varianza*** se obtiene dividiendo entre n la proporción teórica de éxitos multiplicada por su valor complementario. Y su ***forma*** se ajusta a la distribución binomial con parámetros n y π_1. Lo cual significa que es posible utilizar la distribución binomial para conocer las probabilidades asociadas a cada posible valor de los estadísticos n_1 y P_1 en las diferentes muestras de tamaño n.

Pero la distribución binomial no es la única que nos puede ofrecer información sobre las probabilidades asociadas a los estadísticos n_1 y P_1. Sabemos por el *teorema del límite central* que, conforme el número de ensayos n va aumentando[9], la distribución binomial se va pareciendo (aproximando) más y más a la distribución normal (ver, en el capítulo 5, el apartado *Aproximación de la binomial a la normal*). Por tanto, conforme el tamaño muestral va aumentando, la forma de la distribución muestral de n_1 y de P_1 se va aproximando más y más a la distribución normal[10] con los parámetros definidos en [6.6] para n_1 y los definidos en [6.7] para P_1. Ahora bien, si la distribución binomial se aproxima a la normal a medida que el tamaño muestral va aumentando, entonces

$$Z = \frac{n_1 - n\pi_1}{\sigma_{n_1}} = \frac{P_1 - \pi_1}{\sigma_{P_1}} \sim N(0,\ 1) \qquad [6.8]$$

Esto significa que, con tamaños muestrales grandes, es posible utilizar las probabilidades de la distribución normal como una aproximación a las probabilidades de la distribución binomial[11]. La aproximación es tanto mejor cuanto mayor es el tamaño muestral.

Ejemplo. Distribución muestral del estadístico proporción

Este ejemplo muestra cómo obtener la distribución muestral del estadístico P_1 y cómo utilizar las probabilidades de la distribuciones binomial y normal para conocer las probabilidades asociadas a los diferentes valores del estadístico P_1.

---

[9] Con el estadístico *proporción* ocurre lo mismo que con el estadístico *media*: su distribución muestral se parece a la distribución normal con tamaños muestrales relativamente pequeños. Si π_1 no toma valores extremos (es decir, si π_1 toma valores próximos a 0,5), el grado de parecido es muy alto incluso con tamaños muestrales muy pequeños ($n = 5$); cuanto más se aleja de 0,5 el valor de π_1, mayor necesita ser el tamaño muestral para que el parecido entre ambas distribuciones resulte satisfactorio (ver, en el Capítulo 5, el apartado *Aproximación de la distribución binomial a la normal*).

[10] El teorema del límite central no sólo es aplicable al estadístico $\bar{Y}$ sino también al estadístico *suma total*, es decir, a $n\bar{Y}$. En el contexto de la aproximación binomial a la normal esto significa que tanto el estadístico P_1 (que es una media) como el estadístico n_1 (que es un total: $n_1 = nP_1$) tienden a distribuirse normalmente a medida que n va aumentando.

[11] El grado de parecido entre las probabilidades normales y las binomiales es todavía mayor si se aplica una modificación llamada *corrección por continuidad*:

$$Z = \frac{(n_1 \pm 0{,}5) - n\pi_1}{\sigma_{n_1}} = \frac{(P_1 \pm 0{,}5/n) - \pi_1}{\sigma_{P_1}} \qquad [6.9]$$

La variable n_1 = "número de éxitos en n ensayos" es una variable discreta: en 10 ensayos puede haber 0, 1, 2, etc., éxitos, pero no puede haber, por ejemplo, 4,5 éxitos. Al intercambiar las probabilidades normales y las binomiales se está intercambiando una distribución continua por una discreta. Para mejorar el resultado de este intercambio se puede intentar que, de alguna manera, los valores discretos se conviertan en continuos considerando, por ejemplo, que obtener 8 éxitos o más es equivalente a obtener una puntuación de 7,5 o mayor (ver, en el Capítulo 5, el apartado *Aproximación normal a la binomial*).

Supongamos que un sujeto responde a una prueba de rendimiento que consta de 5 preguntas. Supongamos además que cada pregunta tiene 5 alternativas de respuesta de las que sólo una es correcta. Supongamos, por último, que el sujeto responde al azar y que definimos la variable n_1 = «número de aciertos» y su equivalente P_1 = «proporción de aciertos». Estamos interesados en: (1) construir la distribución muestral de los estadísticos n_1 y P_1, y (2) conocer la probabilidad de que el sujeto acierte más de 3 preguntas:

(1) Para construir la distribución muestral de los estadísticos n_1 y P_1 tenemos dos alternativas: obtener todos los posibles resultados que pueden darse (es decir, todas las muestras posibles de resultados) o recurrir a la distribución binomial que, según sabemos, es la distribución teórica a la que se ajusta una variable del tipo "proporción de éxitos en n ensayos independientes de una variable dicotómica"). Lógicamente, utilizaremos la segunda alternativa por ser más breve y sencilla.

El sujeto puede acertar entre ninguna y todas las preguntas. Por tanto, los posibles resultados del estadístico n_1 irán de 0 (ningún acierto) a 5 (todos los posibles); y los posibles resultados del estadístico $P_1 = n_1/n$ irán de 0/5 a 5/5. Tenemos 5 ensayos ($n = 5$) de una variable dicotómica (acierto-error) con probabilidad de acierto constante en cada ensayo ($\pi_1 = 0{,}20$, pues el sujeto responde al azar y sólo una de las 5 alternativas es correcta). Cuando se dan estas circunstancias, sabemos (ver [3.4]) que

$$P(n_1) \;=\; \frac{n!}{n_1!\,(n-n_1)!}\;\pi_1^{n_1}\,(1-\pi_1)^{n-n_1}$$

Por tanto,

$$P(n_1 = 0) \;=\; P(P_1 = 0/5) \;=\; \frac{5!}{0!\,(5-0)!}\,0{,}20^0(1-0{,}20)^{5-0} \;=\; 0{,}3277$$

$$P(n_1 = 1) \;=\; P(P_1 = 1/5) \;=\; \frac{5!}{1!\,(5-1)!}\,0{,}20^1(1-0{,}20)^{5-1} \;=\; 0{,}4096$$

Aplicando esta misma fórmula para 2, 3, 4 y 5 aciertos se obtienen las probabilidades que muestra la Tabla 6.5. Por supuesto, estas probabilidades pueden obtenerse directamente de la Tabla B del apéndice final sin necesidad de efectuar cálculos (debe tenerse en cuenta que las probabilidades que ofrece la Tabla B están acumuladas).

Tabla 6.5. Distribución muestral de n_1 = «número de aciertos» y P_1 = «proporción de aciertos»

n_1	P_1	$f(n_1) = f(P_1)$
0	0/5	0,3277
1	1/5	0,4096
2	2/5	0,2048
3	3/5	0,0512
4	4/5	0,0064
5	5/5	0,0003

(2) La probabilidad de que un sujeto que responde al azar acierte más de tres preguntas es la probabilidad de que acierte 4 o 5. En la distribución muestral del número de aciertos (Tabla 6.5) puede comprobarse que esta probabilidad vale $0{,}0064 + 0{,}0003 = 0{,}0067$.

Consideremos ahora un escenario distinto. Supongamos que un sujeto responde a una prueba de rendimiento que consta de 20 preguntas. Supongamos además que cada pregunta tiene 2 alternativas de respuesta de las que sólo una es correcta. Supongamos, por último, que el sujeto responde al azar y que definimos la variable n_1 = «número de aciertos». Estamos interesados en conocer la probabilidad de que el sujeto acierte más de 14 preguntas.

En este nuevo escenario tenemos 20 ensayos (n= 20) de una variable dicotómica (acierto-error) con probabilidad de acierto constante en cada ensayo (π_1= 0,50, pues el sujeto responde al azar y sólo una de las 2 alternativas es correcta). Cuando se dan estas circunstancias, sabemos que la variable n_1 se distribuye binomialmente. Pero como el tamaño muestral es razonablemente grande y π_1 toma un valor centrado, vamos a utilizar la distribución normal como una aproximación a las probabilidades de la distribución binomial. De acuerdo con [6.8],

$$Z = \frac{n_1 - \mu_{n_1}}{\sigma_{n_1}} = \frac{n_1 - n\pi_1}{\sqrt{n\pi_1(1-\pi_1)}} = \frac{15 - 20(0,50)}{\sqrt{20(0,50)(1-0,50)}} = 2,24 \quad \sim \quad N(0,\,1)$$

(se ha tipificado el valor 15 porque interesa conocer la probabilidad asociada a más de 14 aciertos; es decir, la probabilidad asociada a 15, 16, ..., 20 aciertos). El resultado obtenido indica que la probabilidad de encontrar más de 14 aciertos por azar, o $P(n_1 > 14)$, es igual a la probabilidad de obtener valores Z mayores que 2,24, o $P(Z > 2,24)$. En la tabla de la distribución normal tipificada encontramos,

$$P(Z > 2,24) = 1 - F(2,24) = 1 - 0,9875 = 0,0125$$

Por tanto, la probabilidad de que un sujeto que responde al azar acierte más de 14 preguntas vale 0,0125. Puede comprobarse que esta probabilidad es parecida a la que se obtiene consultando la probabilidad exacta que ofrece la tabla de la distribución binomial[12], donde

$$1 - F(14) = 1 - 0,979 = 0,021$$

Importancia del tamaño muestral

El *error típico* de una distribución muestral (al igual que su cuadrado, la varianza) es un dato de suma importancia en el contexto de la inferencia estadística. Al informar del grado de dispersión de la distribución muestral, también lo está haciendo del grado de precisión del estadístico, es decir, del grado de parecido que cabe esperar encontrar entre un estadístico y su correspondiente parámetro: cuanto menor sea el error típico, menor será la variabilidad del

[12] El grado de parecido entre la probabilidad normal y la probabilidad binomial es todavía mayor si se aplica la corrección por continuidad propuesta en la nota a pie de página número 9 de este mismo capítulo:

$$Z = \frac{n_1 \pm 0,5 - n\pi_1}{\sqrt{n\pi_1(1-\pi_1)}} = \frac{14,5 - 20(0,50)}{\sqrt{20(0,50)(1-0,50)}} = 2,01$$

Nótese que ahora se ha tipificado el valor 14,5. Con esta modificación, la probabilidad de encontrar más de 14 aciertos, o $P(X > 14)$, es igual a la probabilidad de obtener valores Z mayores que 2,01, o $P(Z > 2,01)$. En la tabla de la distribución normal tipificada encontramos: $P(Z > 2,01) = 1 - F(2,01) = 1 - 0,9778 = 0,022$.

estadístico (es decir, más parecidos serán entre sí los posibles valores del estadístico) y, consiguientemente, más parecido habrá entre el valor que finalmente tome en una muestra concreta y el verdadero valor de su correspondiente parámetro.

Y ocurre que el error típico está estrechamente relacionado con el tamaño muestral. Esto puede apreciarse fácilmente tanto en el error típico de la media como en el de la proporción; puesto que en ambos casos el tamaño muestral, n, se encuentra en el denominador (está dividiendo), el error típico será tanto más pequeño cuanto mayor sea n.

En la distribución muestral de $\bar{Y}$, por ejemplo, el error típico es $\sigma_{\bar{Y}} = \sigma_Y / \sqrt{n}$. En esta fórmula es fácil constatar que, a medida que n va aumentando, $\sigma_{\bar{Y}}$ va disminuyendo: con muestras de tamaño $n = 1$, $\sigma_{\bar{Y}}$ es igual a σ_Y; con muestras de tamaño $n = 25$, $\sigma_{\bar{Y}}$ es la quinta parte de σ_Y; con muestras de tamaño $n = 100$, $\sigma_{\bar{Y}}$ es la décima parte de σ_Y; etc. Es claro que, conforme el tamaño muestral va aumentando, el error típico va disminuyendo. Y conforme el error típico va disminuyendo, la variabilidad del estadístico $\bar{Y}$ va siendo menor, lo que implica que *los posibles valores que puede tomar $\bar{Y}$ se parecerán cada vez más a su valor esperado*, que no es otro que la media de la población (μ_Y).

Apéndice 6

Valor esperado y varianza del estadístico *media*

Una muestra aleatoria de tamaño n es una secuencia de n variables aleatorias ($Y_1, Y_2, ..., Y_n$) igualmente distribuidas y, si el muestreo es con reposición, o sin reposición en población infinita, independientes entre sí. Si hacemos:

$$Y = k_1 Y_1 + k_2 Y_2 + \cdots + k_n Y_n$$

entonces la variable combinada Y es también una variable aleatoria distribuida exactamente igual que $Y_1, Y_2, ..., Y_n$, con:

$$E(Y) = \mu_Y = \sum k_i E(Y_i) \qquad \text{y} \qquad V(Y) = \sigma_Y^2 = \sum k_i^2 \sigma_{Y_i}^2$$

Y dado que $Y_1, Y_2, ..., Y_n$ tienen la misma distribución, también tendrán los mismos valores esperados y varianzas:

$$E(Y_1) = E(Y_2) = \cdots = E(Y_n) = \mu_Y \qquad \text{y} \qquad \sigma_{Y_1}^2 = \sigma_{Y_2}^2 = \cdots = \sigma_{Y_n}^2 = \sigma_Y^2$$

Por tanto,

$$E(\bar{Y}) = \mu_{\bar{Y}} = \frac{1}{n} E(Y_1 + Y_2 + \cdots + Y_n) = \frac{1}{n}[E(Y_1) + E(Y_2) + \cdots + E(Y_n)] =$$

$$= \frac{1}{n}(\mu_1 + \mu_2 + \cdots + \mu_n) = \frac{1}{n} n \mu_Y = \mu_Y = \mu_Y \qquad [6.10]$$

$$V(\bar{Y}) = \sigma_{\bar{Y}}^2 = \frac{1}{n^2}\sigma_{Y_1+Y_2+\cdots+Y_n}^2 = \frac{1}{n^2}\sigma_{Y_1}^2 + \sigma_{Y_2}^2 + \cdots + \sigma_{Y_n}^2 = \frac{1}{n^2} n \sigma_Y^2 = \frac{\sigma_Y^2}{n}$$

Distribución muestral del estadístico *varianza*

Hemos estudiado la distribución muestral de dos estadísticos: la proporción y la media. La primera es un ejemplo de distribución muestral discreta basada en una variable categórica; la segunda es un ejemplo de distribución muestral continua basada en una variable cuantitativa. Aunque el concepto de distribución muestral haya podido quedar claro con el estudio de estas dos distribuciones muestrales, todavía nos falta por estudiar otra distribución muestral muy relacionada con el análisis de datos (ver, por ejemplo, Apéndice 9). Nos referimos a la distribución muestral de la *varianza*[13].

Consideremos una población cualquiera y una variable aleatoria Y_i definida en ella. Extraigamos de esa población muestras aleatorias de tamaño n y calculemos S_Y^2 en cada una de ellas. A medida que vayamos extrayendo más y más muestras y calculando S_Y^2 iremos recopilando la información necesaria para conocer la *distribución muestral de la varianza*. Pues bien, esta distribución muestral se encuentra estrechamente relacionada con la distribución ji-cuadrado estudiada en el Apéndice 5. Si la variable Y_i se distribuye normalmente $N(\mu, \sigma)$, entonces

$$\chi^2_{n-1} = \frac{(n-1)\,S_Y^2}{\sigma_Y^2} \quad \rightarrow \quad S_Y^2 = \frac{\sigma_Y^2}{n-1} \qquad [6.11]$$

(ver, por ejemplo, Pardo y San Martín, 1998, págs. 72-73). En efecto, de acuerdo con las ecuaciones [5.13] y [5.14],

$$\chi^2_n = \sum_{i=1}^{n} Z_i^2 = \frac{\sum_i^n (Y_i - \mu)^2}{\sigma_Y^2} \qquad [6.12]$$

El problema de esta ecuación es que raramente se conoce el valor de la media poblacional. No obstante, si la media poblacional μ se sustituye por su valor muestral $\bar{Y}$, la transformación

$$\chi^2_{n-1} = \frac{\sum_i^n (Y_i - \bar{Y})^2}{\sigma_Y^2} \qquad [6.13]$$

todavía se distribuye según χ^2, aunque se pierde un grado de libertad al tener que estimar la media. Y como la varianza muestral se define como

$$S_Y^2 = \sum_i^n (Y_i - \bar{Y})^2/(n-1) \qquad [6.14]$$

se llega de forma directa a las expresiones propuestas en [6.11]. Ahora bien, como $\sigma_Y^2/(n-1)$ es constante para una población y tamaño muestral dados, la distribución muestral de S_Y^2 sólo difiere de χ^2_{n-1} por una constante. Por tanto, la distribución muestral de la varianza tiene la forma de una distribución ji-cuadrado con $n-1$ grados de libertad. Y como el valor esperado de una variable ji-cuadrado son sus grados de libertad y la varianza de una variable ji-cuadrado es el doble de sus grados de libertad (ver ecuación [5.16]), a partir de [6.11] se obtiene

$$\begin{aligned} E(S_Y^2) &= E[\chi_{n-1}\,\sigma_Y^2/(n-1)] = (n-1)\,\sigma_Y^2/(n-1) = \sigma_Y^2 \\ V(S_Y^2) &= V[\chi_{n-1}\,\sigma_Y^2/(n-1)] = 2(n-1)\,\sigma_Y^4/(n-1)^2 = 2\sigma_Y^4/(n-1) \end{aligned} \qquad [6.15]$$

---

[13] Al iniciarse en el análisis de datos suele sorprender oír hablar de la media de la *media* o de la varianza de la *varianza*. Debe tenerse en cuenta que tanto el estadístico *media* como el estadístico *varianza* son variables (todos los estadísticos son variables). Y, como variables que son, tienen su propia distribución, la cual a su vez tiene media y varianza.

El método Monte Carlo

El método **Monte Carlo** es un método de simulación muy útil para extraer muestras aleatorias de poblaciones concretas y, en lo que aquí más nos interesa, para obtener las distribuciones muestrales de algunos estadísticos cuando la situación resulta matemáticamente intratable. Los párrafos que siguen no pretenden ofrecer aquí una explicación exhaustiva del método Monte Carlo y de todas sus posibilidades pues eso excedería ampliamente las pretensiones de este apartado, pero sí se presentan unos sencillos ejemplos que pueden ayudar a comprender la utilidad del muestreo simulado.

Supongamos que deseamos estudiar algunos aspectos relacionados con una variable dicotómica (es decir, una variable que sólo puede tomar dos valores: acierto-error, presencia-ausencia, verdadero-falso, etc.). Llamemos *éxito* y *fracaso* de forma genérica a cada uno de esos dos valores. Supongamos, además, que $\pi_{\text{éxito}} = 0{,}70$ y $\pi_{\text{fracaso}} = 0{,}30$ son las probabilidades poblacionales asociadas a cada uno de los dos valores de esa variable dicotómica. Supongamos, por último, que la población en la que deseamos estudiar esa variable es infinita o tan grande que a todos los efectos puede ser considerada infinita.

De esa población extraemos una muestra aleatoria de tamaño 100 en la que definimos la variable n_1 = «número de éxitos» (siendo *éxito* uno cualquiera de los dos niveles de la variable definida en la población). Con un ordenador o, incluso, con una calculadora de bolsillo, se pueden generar 100 números aleatorios u_i entre 0 y 1 (cualquiera de estos números u_i son valores de una distribución uniforme: todos ellos tienen la misma probabilidad de aparecer). Si el número u_i es menor o igual que 0,70, consideraremos que hemos extraído un elemento perteneciente a la categoría *éxito*; si el número u_i es mayor que 0,70, consideraremos que hemos extraído un elemento perteneciente a la categoría *fracaso*. Estos 100 números u_i constituyen una muestra aleatoria procedente de una población binomial (ver Capítulo 3) con parámetros $n = 100$ y $\pi_{\text{éxito}} = 0{,}70$.

Pero, ¿para qué sirve extraer una muestra si ya se conocen las características de la población? Según veremos, las herramientas inferenciales (la estimación y el contraste) se basan en el concepto de *distribución muestral*. La distribución muestral de un estadístico es la distribución de probabilidad que se obtiene al calcular ese estadístico en todas las muestras de tamaño n que es posible extraer de una determinada población. Hemos visto que, con poblaciones pequeñas, no es demasiado complicado obtener una distribución muestral. Sin embargo, con poblaciones muy grandes el proceso de obtención de una distribución muestral puede resultar muy largo y tedioso. Y, por supuesto, con poblaciones infinitas no resulta posible extraer las infinitas muestras de cualquier tamaño que sería posible definir. Para abordar este tipo de situaciones existen procedimientos analíticos que permiten obtener la distribución muestral de algunos estadísticos (esto es lo que hemos hecho con los estadísticos $\bar{Y}$ y P). Además, hay situaciones en las que los procedimientos analíticos no resultan útiles porque, o bien no son aplicables (no hay procedimientos matemáticos capaces de ofrecer una solución), o bien resultan demasiado engorrosos y, por tanto, poco prácticos (la situación es matemáticamente intratable debido a su complejidad). Es justamente en estos casos donde puede utilizarse un método de simulación como el método Monte Carlo para generar, no infinitas muestras, por supuesto, pero sí un número de ellas lo bastante grande como para obtener una distribución muestral lo bastante precisa.

Siguiendo con nuestro ejemplo, sabemos que la variable n_1 = «número de éxitos» definida más arriba se distribuye binomialmente con valor esperado $n\pi_1$ y varianza $n\pi_1(1 - \pi_1)$. Pero si no hubiera forma de conocer estos valores por procedimientos matemáticos, podríamos generar, por ejemplo, 10.000 muestras como la referida más arriba y calcular en cada una de ellas el valor de n_1. Tendríamos así 10.000 valores n_1 que estarían ofreciendo una información bastante precisa de su verdadero valor esperado, de su varianza y de la forma de su distribución. Y eso permitiría conocer con bastante precisión la distribución muestral de n_1.

Lo mismo que se hace con una variable distribuida binomialmente también podría hacerse con otro tipo de variables. Consideremos el caso de una variable cuantitativa cualquiera Y cuyas frecuencias relativas $[f(Y)]$ y relativas acumuladas $[F(Y)]$, en la población, son las que recoge la Tabla 6.6. Para extraer de esa población una muestra aleatoria de tamaño $n = 100$ se puede proceder de la siguiente manera: se genera un número aleatorio u_i entre 0 y 1. Si u_i es igual o menor que 0,23, se considera que se ha

obtenido un elemento muestral $Y_i = 0$; si u_i es mayor que 0,23 y menor o igual que 0,57, se considera que se ha obtenido un elemento muestral $Y_i = 1$; si u_i es mayor que 0,57 y menor o igual que 0,80, se considera que se ha obtenido un elemento muestral $Y_i = 2$; etc. Tras generar 100 números aleatorios u_i tendremos una muestra aleatoria de tamaño $n = 100$. Repitiendo el proceso, por ejemplo, 10.000 veces (tarea bastante simple de realizar con un ordenador) tendremos 10.000 muestras aleatorias de tamaño $n = 100$. Si, ahora, en cada una de esas muestras se calcula el estadístico $\bar{Y}$, será posible construir de forma bastante precisa la distribución muestral del estadístico $\bar{Y}$. Vemos, con estos sencillos ejemplos, que el método Monte Carlo permite generar cualquier número de muestras aleatorias, y de cualquier tamaño, siempre que se conozca la forma de la distribución poblacional.

Tabla 6.6. Distribución de probabilidad de la variable *Y*

Y	$f(Y)$	$F(Y)$
0	0,23	0,23
1	0,34	0,57
2	0,23	0,80
3	0,12	0,92
4	0,06	0,98
5	0,02	1,00

Por supuesto, cuanto más complejas son las distribuciones poblacionales, más complicado resulta generar muestras aleatorias mediante simulación. Los ejemplos que acabamos de presentar se refieren a distribuciones bastante simples (la binomial y la multinomial) y por esta razón la obtención de muestras aleatorias resulta también bastante sencilla. Con distribuciones más complejas (como, por ejemplo, la distribución normal), el método de extracción se vuelve algo más complicado y, sobre todo, menos intuitivo. En el caso concreto de la distribución normal existen diferentes procedimientos que permiten obtener muestras aleatorias de distribuciones $N(\mu, \sigma)$. Algunos de esos procedimientos se basan, al igual que en los ejemplos propuestos, en números aleatorios distribuidos de forma uniforme en el rango (0, 1); tal es el caso, por ejemplo, de la muy conocida y utilizada técnica *Box-Muller* (ver, por ejemplo, Lewis y Orav, 1989, págs. 45-47). Otros, más sofisticados, se basan en números aleatorios no distribuidos uniformemente.

No es nuestro propósito detallar aquí esos procedimientos, sino sólo mencionar que existen y, lo que es más importante, señalar que, en la mayor parte de los programas informáticos de análisis estadístico (y, desde luego, en el SPSS), existen diferentes rutinas que permiten extraer muestras aleatorias de prácticamente cualquier tipo de distribución conocida.

Ejercicios

6.1. A partir de una población formada por los siguientes $N = 5$ elementos $Y = \{0, 2, 4, 6 \text{ y } 8\}$, obtener:

a. La media y la varianza de la población (μ_Y y σ_Y^2).

b. El número de muestras aleatorias de tamaño $n = 2$ que es posible extraer, con reposición, de esa población.

c. La distribución muestral de la media $\bar{Y}$.

6.2. Con los datos del ejercicio 6.1:

a. Calcular la media y la varianza de la distribución muestral de la media.
b. Determinar la relación existente entre los estadísticos y los parámetros.

6.3. Un test está formado por 3 preguntas con dos alternativas de respuesta cada una: V = «verdadero», F = «falso». Si el sujeto responde al azar y definimos la variable n_V = «número de respuestas V»:

a. ¿Cuál es la función de probabilidad de la variable n_V?
b. Calcular el valor esperado y la varianza de n_V.

6.4. En una determinada empresa se sabe que el 40% de los empleados están a favor de reducir la jornada laboral aunque ello implique una rebaja en el salario. Se seleccionan aleatoriamente 8 empleados y se les pregunta si están a favor o en contra de la reducción de jornada:

a. Construir la distribución muestral del estadístico n_F = «número de empleados a favor de la reducción de jornada».
b. Calcular el valor esperado y la varianza del estadístico n_F.

6.5. Sabemos que la población de estudiantes de enseñanza primaria se distribuye $N(8, 3)$ en una prueba de comprensión lectora (Y). Al seleccionar de esa población una muestra aleatoria de 9 estudiantes y aplicar la prueba de comprensión lectora, ¿cuál es la probabilidad de obtener una media comprendida entre 7 y 10?

6.6. Las puntuaciones en una prueba de madurez (Y) se distribuyen normalmente con media 5 y varianza 9 en la población de estudiantes de enseñanza primaria. Generalmente, los grupos cuyo promedio en Y no alcanza el percentil 20 terminan necesitando el apoyo de un experto en educación especial. En un grupo de 25 niños se ha obtenido $\bar{Y} = 4{,}4$.

a. ¿Decidirá el profesor de ese grupo pedir la colaboración de un experto en educación especial? Es decir, ¿se encuentra el promedio 4,4 por debajo del percentil 20?
b. ¿Se llegaría a la misma conclusión si el promedio $\bar{Y} = 4{,}4$ se hubiera obtenido en un grupo de 16 estudiantes?

6.7. Supongamos que la población de estudiantes que acceden a la universidad distribuye normalmente con $\mu_Y = 40$ en una prueba de aptitud para las matemáticas. Al seleccionar una muestra aleatoria de 25 estudiantes de esa población se ha obtenido una media de 43 y una desviación típica de 17,5. ¿Cuál es la probabilidad de encontrar medias como la obtenida o mayores?

6.8. El nivel de desarrollo cognitivo ha sido repetidamente identificado como un factor de importancia fundamental para explicar el rendimiento académico en los primeros cursos de enseñanza primaria. En una escala de desarrollo cognitivo (Y), los niños de 6 años alcanzan una puntuación media de 12. Una muestra aleatoria de 45 niños ha obtenido en esa escala de desarrollo una media de 14 y una varianza de 27. ¿Cuál es la probabilidad de obtener medias iguales o mayores que la obtenida?

6.9. Considerando que el clima de clase influye decisivamente sobre el rendimiento académico en los primeros cursos de enseñanza secundaria, un educador decide aplicar un programa de

modificación de conducta a los alumnos de un grupo de 14 años si éstos superan, en promedio, el percentil 25 en una escala (Y) diseñada para evaluar el clima de clase (las puntuaciones bajas indican *buen clima* y las altas *mal clima*). El educador sabe que, en la población de estudiantes de 14 años, las puntuaciones de esa escala se distribuyen normalmente con media 50 y desviación típica 9. Si aplica la escala a los 36 estudiantes del mencionado grupo, ¿qué media, como mínimo, deben obtener esos 36 estudiantes para que el educador decida aplicar el programa?

6.10. Las puntuaciones de una escala de depresión (Y) se distribuyen normalmente en la población de adultos. El distribuidor de la escala afirma que el 25% de los sujetos obtiene puntuaciones menores que 20 y otro 25% puntuaciones mayores que 50. Con esta información:

a. Calcular la media y la varianza de la población.

b. Calcular la probabilidad de obtener en esa escala una media igual o mayor que 40 en una muestra aleatoria de 25 sujetos.

6.11. Siendo $\overline{Y}$ la media aritmética de una muestra aleatoria de tamaño 9 extraída de una población $N(1, 3)$, calcular el valor de $\overline{Y}$ sabiendo que la probabilidad de obtener medias como ésa o mayores vale 0,25.

6.12. ¿Cuál debe ser el tamaño de una muestra aleatoria extraída de una población en la que Y se distribuye $N(40, 10)$ para que valga 0,99 la probabilidad de que la media de dicha muestra sea menor que 42?

6.13. La variable Y se distribuye $N(10, 2)$ en una determinada población. Al extraer dos muestras aleatorias de tamaños $n_1 = 9$ y $n_2 = 16$:

a. ¿$P(\overline{Y}_1 > 10) = P(\overline{Y}_2 > 10)$? ¿Por qué?

b. ¿$P(\overline{Y}_1 > 20) = P(\overline{Y}_2 > 20)$? ¿Por qué?

6.14. La variable aleatoria Y se distribuye normalmente $N(30, 10)$. A continuación se ofrece la función de distribución (probabilidades acumuladas) de algunos de sus valores. Completar la tabla teniendo en cuenta que, en lo relativo a la distribución muestral de la media se está trabajando con muestras de tamaño 25.

Y	0	10	20	30	40	50	60
Z	( )	( )	( )	( )	( )	( )	( )
$F(Z)$	( )	( )	( )	( )	0,840	0,977	0,999
$\overline{Y}$	( )	( )	( )	( )	( )	( )	( )

6.15. Supongamos que, de una población con dos valores equiprobables, $Y = \{0, 1\}$, se extrae con reposición una muestra aleatoria de tamaño $n = 2$. ¿Cuál es la probabilidad de que la media muestral valga 1?

6.16. En un experimento sobre agudeza visual se ha presentado a un sujeto 50 pares de estímulos luminosos para comprobar si era capaz de percibir la diferencia en intensidad entre los dos

estímulos de cada par. El sujeto debía pulsar un botón rojo cuando creía que la intensidad luminosa de los estímulos era distinta y un botón verde cuando creía que la intensidad luminosa de los estímulos era la misma. Si consideramos que el sujeto ha estado pulsando los botones al azar, ¿cuál es la probabilidad de encontrar más de 30 aciertos?

6.17. Los datos recogidos en los últimos años indican que el 65% de los estudiantes universitarios muestra una actitud favorable hacia la eutanasia.

a. Calcular el valor esperado y la varianza de la variable dicotómica X = «actitud hacia la eutanasia» (a favor, en contra).

b. Calcular el valor esperado y la varianza de la distribución muestral de n_F = «número de universitarios que se muestran a favor de la eutanasia» al extraer muestras aleatorias de 20 universitarios.

c. ¿Cuál es la probabilidad de encontrar, en la muestra de 20 universitarios, más de 15 con actitud favorable?

6.18. Al parecer, los pacientes aquejados de neurosis depresiva se recuperan espontáneamente, es decir, sin necesidad de tratamiento, en el 30% de los casos (transcurridos 3 meses desde el inicio del trastorno). En la lista de espera de un hospital hay 20 pacientes diagnosticados con neurosis depresiva que no recibirán tratamiento antes de 3 meses. ¿Cuál es la probabilidad de que, transcurridos 3 meses, más de 8 de esos 20 pacientes no necesiten tratamiento?

6.19. Consideremos un test formado por 10 preguntas, cada uno de las cuales consta de 4 alternativas de respuesta con sólo una correcta. Si una muestra de 100 sujetos responde al azar a las 10 preguntas, ¿cuántos sujetos cabe esperar que acierten al menos 5 preguntas?

6.20. En una distribución muestral formada a partir de las medias obtenidas con muestras de tamaño $n = 49$, a la media $\overline{Y} = 76$ le corresponde una puntuación típica $Z = 2$. Si el error típico de esa distribución muestral vale 3:

a. ¿Cuál es el valor de la media poblacional?

b. ¿Cuál es el valor de la desviación típica poblacional?

6.21. En un examen tipo test con n preguntas, cada una de las cuales tiene k alternativas de respuesta de las que sólo una es correcta, sabemos que la variable n_A = «número de aciertos» tiene un valor esperado de 10 y una varianza de 8:

a. ¿Cuántas preguntas tiene el examen?

b. ¿Cuántas alternativas de respuesta tiene cada pregunta?

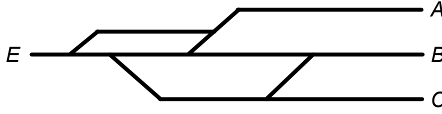
6.22. La variable aleatoria Y se distribuye $N(\mu_Y, \sigma_Y)$ en una determinada población. Se extraen infinitas muestras aleatorias de tamaño n de esa población y se calcula en todas ellas la media aritmética $\overline{Y}$.

a. Si $n \geq 2$, ¿puede afirmarse que σ_Y siempre es mayor que $\sigma_{\overline{Y}}$? ¿Por qué?

b. Siendo k un valor cualquiera, ¿puede afirmarse que $P(\overline{Y} \geq k)$ es menor que $P(Y \geq k)$?

6.23. En un estudio sobre aprendizaje con ratas, un psicólogo ha utilizado un laberinto como el que muestra la figura que se ofrece a continuación. El laberinto tiene una entrada (E) y 3 salidas

diferentes (*A*, *B* y *C*). Al final de cada salida hay una recompensa en cantidad de bolas de comida: 1 bola en la salida *A*, 2 en la *B* y 3 en la *C*.



Asumiendo que las ratas siempre caminan hacia adelante, es decir, sin retroceder:

a. Obtener la distribución de probabilidad de la variable X = «número de bolas de comida conseguidas al recorrer el laberinto».

b. Calcular el valor esperado y la varianza de la variable X.

6.24. Supongamos ahora que el laberinto lo recorren, independientemente, dos ratas. Obtener el valor esperado y la varianza de la variable Y = «número de bolas de comida conseguidas por las dos ratas al recorrer el laberinto».

Soluciones

6.1. *a*. $\mu_Y = E(Y) = \sum Y_i f(Y_i) = (0+2+4+6+8)/5 = 4.$

$\sigma_Y^2 = V(Y) = E(Y^2) - [E(Y)]^2 = 24 - 4^2 = 8.$

$E(Y^2) = \sum Y_i^2 f(Y_i) = (0^2+2^2+4^2+6^2+8^2)/5 = 120/5 = 24.$

b. Número de muestras con reposición: $N^n = 5^2 = 25.$

c.

Muestras	$\bar{Y}$	$f(\bar{Y})$
(0,0)	0	1/25
(0,2) (2,0)	1	2/25
(0,4) (2,2) (4,0)	2	3/25
(0,6) (2,4) (4,2) (6,0)	3	4/25
(0,8) (2,6) (4,4) (6,2) (8,0)	4	5/25
(2,8) (4,6) (6,4) (8,2)	5	4/25
(2,4) (4,2) (6,6)	6	3/25
(6,8) (8,6)	7	2/25
(8,8)	8	1/25

6.2. *a*. $\mu_{\bar{Y}} = E(\bar{Y}) = \sum \bar{Y}_i f(\bar{Y}_i) = 0\,(1/25) + 1\,(2/25) + \cdots + 8\,(1/25) = 4.$

$\sigma_{\bar{Y}}^2 = V(\bar{Y}) = E(\bar{Y}^2) - [E(\bar{Y})]^2 = 20 - 4^2 = 4.$

$E(\bar{Y}^2) = \sum \bar{Y}_i^2 f(\bar{Y}_i) = 0^2(1/25) + 1^2(2/25) + \cdots + 8^2(1/25) = 20.$

b. $\mu_Y = \mu_{\bar{Y}} = 4; \quad \sigma_{\bar{Y}}^2 = \sigma_Y^2/n = 8/2 = 4.$

6.3. *a*. Distribución muestral del número o proporción de aciertos (= respuestas *V*).

n_V	*Muestras*	$f(n_V)$
0	FFF	1/8
1	VFF, FVF, FFV	3/8
2	VVF, VFV, FVV	3/8
3	VVV	1/8

b. $E(n_V) \;=\; \sum n_V f(n_V) \;=\; 0(1/8) + 1(3/8) + 2(3/8) + 3(1/8) \;=\; 1{,}5.$

$V(n_V) \;=\; E(n_V^2) - [E(n_V)]^2 \;=\; 3 - 1{,}52 \;=\; 0{,}75\,.$

$E(n_V^2) \;=\; \sum n_V^2 f(n_V) \;=\; 0^2(1/8) + 1^2(3/8) + 2^2(3/8) + 3^2(1/8) \;=\; 24/8 \;=\; 3.$

6.4. *a*. Los posibles valores de n_F van de 0 (nadie a favor) a 8 (todos a favor). Utilizando la distribución binomial con $n = 8$ y $\pi_F = 0{,}40$ se obtienen los resultados que se ofrecen a continuación (si se utiliza la Tabla B del Apéndice final, debe tenerse en cuenta que las probabilidades de la tabla están acumuladas):

n_F	$f(n_F)$
0	0,017
1	0,089
2	0,209
3	0,279
4	0,232
5	0,124
6	0,041
7	0,010
8	0,000

b. $E(n_F) \;=\; n\pi_F \;=\; 8\,(0{,}40) \;=\; 3.2.$

$V(n_F) \;=\; n\pi_F\,(1 - \pi_F) \;=\; 8\,(0{,}40)\,(1 - 0{,}40) \;=\; 1.92.$

6.5. Distribución muestral de la media con $\mu_Y = 8$, $\sigma_Y = 3$ y $n = 9$.

$P(7 \leq \bar{Y} \leq 10) \;=\; P(Z_{(7)} \leq Z \leq Z_{(10)}) \;=\; F(Z_{(10)} - F(Z_{(7)}).$

$$Z_{(\bar{Y})} \;=\; \frac{\bar{Y} - \mu_{\bar{Y}}}{\sigma_Y/\sqrt{n}} \quad\rightarrow\quad Z_{(7)} \;=\; \frac{7-8}{3/\sqrt{9}} \;=\; -1 \quad\rightarrow\quad Z_{(10)} \;=\; \frac{10-8}{3/\sqrt{9}} \;=\; 2.$$

Por tanto, $P(7 \leq \bar{Y} \leq 10) \;=\; P(-1 \leq Z \leq 2) \;=\; F(2) - F(-1) \;=\; 0{,}9772 - 0{,}1587 \;=\; 0{,}8185.$

6.6. Distribución muestral de la media con $\mu_Y = 5$ y $\sigma_Y = 3$.

a. Con $n = 25$: $P(\bar{Y} \leq 4{,}4) \;=\; P(Z \leq \dfrac{4{,}4-5}{3/\sqrt{25}}) \;=\; P(Z \leq -1) \;=\; 0{,}1587.$

Con $n = 25$, a la media 4,4 le corresponde el percentil 15,87, el cual se encuentra por debajo del percentil 20 establecido como criterio para pedir apoyo. Por tanto, el profesor del grupo decidirá pedir la colaboración de un experto en educación especial.

b. Con $n = 16$: $P(\overline{Y} \leq 4{,}4) = P(Z \leq \frac{4{,}4 - 5}{3/\sqrt{16}}) = P(Z \leq -0{,}8) = 0{,}2119.$

Con $n = 16$, a la media 4,4 le corresponde el percentil 21,19, el cual se encuentra por encima del percentil 20. En consecuencia, el profesor del grupo decidirá no pedir la colaboración del experto.

6.7. Distribución muestral de la media con $\mu_Y = 40$, σ_Y desconocida y $n = 25$.
Como desconocemos σ, utilizamos la distribución *t* con $n - 1 = 24$ grados de libertad.

$$P(\overline{Y} \geq 43) = P(T \geq T_{(43)}) = 1 - F(T_{(43)}).$$

$$T_{(\overline{Y})} = \frac{\overline{Y} - \mu_{\overline{Y}}}{S_Y/\sqrt{n}} \quad \rightarrow \quad T_{(7)} = \frac{43 - 40}{17{,}5/\sqrt{25}} = 0{,}857.$$

$$P(\overline{Y} \geq 43) = P(T \geq 0{,}857) = 1 - F(0{,}857) = 1 - 0{,}80 = 0{,}20.$$

6.8. Distribución muestral de la media con $\mu_Y = 12$, σ_Y desconocida y $n = 45$.
Como desconocemos σ_Y, utilizamos la distribución *t* con $n - 1 = 44$ grados de libertad.

$$P(\overline{Y} \geq 14) = P(T \geq T_{(14)}) = 1 - F(T_{(14)}).$$

$$T_{(14)} = \frac{14 - 12}{\sqrt{27/45}} = 2{,}582.$$

$$P(\overline{Y} \geq 14) = P(T \geq 2{,}582) < 0{,}01.$$

La tabla de la distribución *t* no incluye la probabilidad exacta asociada al valor 2,582, pero sabemos que es menor que 0,01.

6.9. Distribución muestral de la media con $\mu_Y = 50$, $\sigma_Y = 9$ y $n = 36$.
Buscamos la media concreta que corresponde al percentil 25 en la distribución muestral de la media con $\mu_{\overline{Y}} = 50$ y $\sigma_{\overline{Y}} = 9/6$. Sabemos que en la distribución normal estandarizada $N(0, 1)$, el percentil 25 vale $-0{,}67$, pues $P(Z \leq -0{,}67) = 0{,}2514 = 0{,}25$. Por tanto,

$$-0{,}67 = \frac{\overline{Y} - 50}{9/\sqrt{36}} \quad \rightarrow \quad \overline{Y} = -0{,}67 \frac{9}{\sqrt{36}} + 50 = 48{,}995.$$

Es decir, al percentil 25 le corresponde una media de 48,995. En consecuencia, ése es el promedio que debe obtener el grupo, como mínimo, para que el educador decida aplicar el programa.

6.10 *a*. La distribución normal es simétrica. Y sabemos que la puntuación 20 deja por debajo de sí un área idéntica a la que deja la puntuación 50 por encima. Esto significa que las puntuaciones 20 y 50 se encuentran a la misma distancia de la media. Por tanto, $\mu_Y = (20 + 50)/2 = 35$. Para calcular la varianza, sabemos que $Z_{(50)} = Z_{0,75} = 0{,}67$. Por tanto,

$$0{,}67 = \frac{50 - 35}{\sigma_Y} \quad \rightarrow \quad \sigma_Y = \frac{50 - 35}{0{,}67} = 22{,}39.$$

b. Distribución muestral de la media con $\mu = 35$, $\sigma = 22{,}39$ y $n = 25$.

$$P(\bar{Y} \geq 40) \;=\; P(Z \geq \frac{40-35}{22{,}39/\sqrt{25}}) \;=\; P(Z \geq 1{,}12) \;=\; 1 - 0{,}8686 = \; 0{,}1314.$$

6.11. Distribución muestral de la media con $\mu_Y = 1$, $\sigma_Y = 3$ y $n = 9$.
Sabemos que $Z_{\bar{Y}} = Z_{0,75} = 0{,}67$. Por tanto,

$$0{,}67 \;=\; \frac{\bar{Y}-1}{3/\sqrt{9}} \quad \rightarrow \quad \bar{Y} \;=\; 0{,}67\,(3/\sqrt{9})+1 \;=\; 0{,}67\,(1)+1 \;=\; 1{,}67.$$

6.12. Distribución muestral de la media con $\mu_Y = 40$ y $\sigma_Y = 10$.
Sabemos que $Z_{(42)} = Z_{0,99} = 2{,}326$. Por tanto,

$$2{,}326 \;=\; \frac{42-40}{10/\sqrt{n}} \quad \rightarrow \quad \sqrt{n} \;=\; \frac{2{,}326\,(10)}{42-40} \quad \rightarrow \quad n \;=\; \left[\frac{2{,}326\,(10)}{42-40}\right]^2 \;=\; 135.$$

6.13. Distribución muestral de la media con $\mu_Y = 10$ y $\sigma_Y = 2$.

a. Sí . El valor $\bar{Y} = 10$ es igual a la media de la población y, por tanto, igual a la media de la distribución muestral de la media. En consecuencia, su tipificación valdrá cero independientemente del tamaño muestral. Y a iguales puntuaciones típicas, igual probabilidad asociada.

b. No. El valor $\bar{Y} = 20$ es mayor que la media de la población. Por tanto, su distancia a la media, una vez tipificado, será tanto menor cuanto mayor sea el tamaño muestral (el cual está dividiendo al tipificar). En consecuencia, $P(\bar{Y}_1 > 20) < P(\bar{Y}_2 > 20)$.

6.14. Distribución normal con $\mu_Y = 30$ y $\sigma_Y = 10$. Y distribución muestral de la media con $n = 25$. A la media le corresponde una puntuación $Z = 0$. Y a cada aumento o disminución de 10 puntos en Y (valor de la desviación típica) le corresponde un aumento o disminución de un punto en Z. Con esta información se pueden obtener las puntuaciones Z. También sabemos que la distribución normal es simétrica. Y esto permite obtener las probabilidades acumuladas $F(Z)$. Por último, el error típico de la distribución muestral vale $10/5 = 2$, lo cual significa que cada distancia de 10 puntos en Y se corresponde con una distancia de 2 puntos en $\bar{Y}$. Esto permite calcular los valores de $\bar{Y}$.

Y	0	10	20	30	40	50	60
Z	−3	−2	−1	0	1	2	3
$F(Z)$	0,001	0,023	0,16	0,5	0,84	0,977	0,999
$\bar{Y}$	24	26	28	30	32	34	36

6.15. Distribución muestral de la media.

Muestras	$\bar{Y}$	$f(\bar{Y})$
(0,0)	0,5	1/4
(0,1) (1,0)	0,5	2/4
(1,1)	1,0	1/4

Por tanto, la probabilidad de que la media valga 1 es 1/4 = 0,25.

6.16. Distribución muestral del número o proporción de aciertos, con $n = 50$ y $\pi_A = 0{,}50$ (A = acierto). Como el tamaño muestral es grande (> 20), utilizamos la aproximación de la binomial a la normal:

$$E(n_A) = n\pi_A = 50\,(0{,}50) = 25.$$
$$V(n_A) = n\pi_A(1-\pi_A) = 50\,(0{,}50)\,(1-0{,}50) = 12{,}5.$$
$$P(n_A > 30) = P(Z \geq \frac{31-25}{\sqrt{12{,}5}}) = P(Z \geq 1{,}70) = 1 - 0{,}9554 = 0{,}0446.$$

6.17. Población dicotómica (1, 0) y distribución muestral del número o proporción de aciertos, con $n = 20$ y $\pi_F = 0{,}65$ (F = «actitud favorable»).

a. $E(X) = \sum X_i\, f(X_i) = 0\,(0{,}35) + 1(0{,}65) = 0{,}65.$

$V(X) = E(X^2) - [E(X)]^2 = 0{,}65 - 0{,}65^2 = 0{,}2275.$

$E(X^2) = \sum X_i^2\, f(X_i) = 0^2(0{,}35) + 1^2(0{,}65) = 0{,}65.$

b. $E(n_F) = n\pi_F = 20\,(0{,}65) = 13.$

$V(n_F) = n\pi_F(1-\pi_F) = 20\,(0{,}65)\,(1-0{,}65) = 4{,}55.$

c. $P(n_F > 15) = P(Z \geq \frac{16-13}{\sqrt{4{,}55}}) = P(Z \geq 1{,}41) = 1 - 0{,}9207 = 0{,}0793.$

6.18. Distribución muestral del número de recuperaciones, con $n = 20$ y $\pi_R = 0{,}30$ (R = «recuperación»).

$$E(n_R) = n\pi_R = 20\,(0{,}30) = 6.$$
$$V(n_R) = n\pi_R(1-\pi_R) = 20\,(0{,}30)\,(1-0{,}30) = 4{,}2.$$
$$P(n_R > 8) = P(Z \geq \frac{9-6}{\sqrt{4{,}2}}) = P(Z \geq 1{,}46) = 1 - 0{,}9279 = 0{,}0721.$$

6.19. Distribución muestral del número de aciertos, con $n = 10$ y $\pi_A = 0{,}25$ (A = «acierto»).

$$E(n_A) = n\pi_A = 10\,(0{,}25) = 2{,}5.$$
$$V(n_A) = n\pi_A(1-\pi_A) = 10\,(0{,}25)\,(1-0{,}25) = 1{,}875.$$
$$P(n_A > 4) = P(Z \geq \frac{5-2{,}5}{\sqrt{1{,}875}}) = P(Z \geq 1{,}83) = 1 - 0{,}9664 = 0{,}0336.$$

Ésta es la probabilidad de que un sujeto acierte al menos 5 preguntas. Si responden al azar 100 sujetos cabe esperar que $100\,(0{,}0336) = 3{,}36$ (al menos 3 sujetos) acierten al menos 5 preguntas.

6.20. Distribución muestral de la media con μ_Y y σ_Y desconocidas, $n = 49$ y $\sigma_{\bar{Y}} = 3$.

$$Z_{(\bar{Y})} = \frac{\bar{Y} - \mu_{\bar{Y}}}{\sigma_Y/\sqrt{n}} \quad \rightarrow \quad 2 = \frac{76 - \mu_Y}{3} \quad \rightarrow \quad \mu_Y = 76 - 2\,(3) = 70.$$

$$\sigma_{\bar{Y}} = \sigma_Y/\sqrt{n} \quad \rightarrow \quad 3 = \sigma_Y/\sqrt{49} \quad \rightarrow \quad \sigma_Y = 3/\sqrt{49} = 21.$$

6.21. Distribución muestral del número de aciertos, con $n\pi_A = 10$ y $n\pi_A(1-\pi_A) = 8$.

$$\pi_A = \frac{10}{n} \quad \rightarrow \quad 8 = n\frac{10}{n}(1 - \frac{10}{n}) = 10 - \frac{100}{n} \quad \rightarrow \quad n = \frac{100}{10-8} = 50.$$

Por tanto, $\pi_A = 10/50 = 0{,}20$. Y, como la probabilidad de acierto es igual a 1 dividido entre el número de alternativas, tendremos: $0{,}20 = 1/k$; de donde $k = 1/0{,}20 = 5$.

6.22. Distribución muestral de la media con μ y σ desconocidas.

a. Sí, pues $\sigma_{\bar{Y}} = \sigma_Y / \sqrt{n}$.

b. No. Cuando $k = \mu_Y$, $P(\bar{Y} \geq k) = P(Y \geq k)$, pues μ_Y es el centro tanto de la distribución de Y como de la distribución $\bar{Y}$.

Cuando $k < \mu_Y$ tampoco se verifica $P(\bar{Y} \geq k) < P(Y \geq k)$, pues al ser la distribución de Y más dispersa que la de $\bar{Y}$, la proporción de área que queda por encima de k es mayor en la distribución de $\bar{Y}$ que en la de Y.

Cuando $k > \mu_Y$ si se verifica $P(\bar{Y} \geq k) < P(Y \geq k)$, pues al ser la distribución de Y más dispersa que la distribución de $\bar{Y}$, la proporción de área que queda por encima de k es mayor en la distribución de Y que en la de $\bar{Y}$.

6.23. *a*. Teniendo en cuenta los distintos caminos por los que puede llegarse a cada salida y el número de bolas (X) de comida que hay en cada salida y asumiendo que, en cada cruce, la probabilidad de optar por cualquiera de los dos caminos es la misma, tenemos:

A la salida *A* (una bola de comida) se puede llegar por dos caminos:

$P(A_1) = 0{,}5.$
$P(A_2) = 0{,}5\,(0{,}5)\,(0{,}5) = 0{,}125.$

A la salida *B* (dos bolas de comida) se puede llegar por dos caminos:

$P(B_1) = 0{,}5\,(0{,}5)\,(0{,}5) = 0{,}125.$
$P(B_2) = 0{,}5\,(0{,}5)\,(0{,}5) = 0{,}125.$

A la salida *C* (tres bolas de comida) se puede llegar por un sólo camino:

$P(C) = 0{,}5\,(0{,}5)\,(0{,}5) = 0{,}125.$

Por tanto,

X	*Salidas*	*Probabilidades*	$f(X)$
1	A_1, A_2	0,5 + 0,125	0,625
2	B_1, B_2	0,125 + 0,125	0,250
3	C	0,125	0,125

b. $E(X) = \sum X_i\, f(X_i) = 1\,(0{,}625) + 2\,(0{,}250) + 3\,(0{,}125) = 1{,}50.$

$V(X) = E(X^2) - [E(X)]^2 = 2{,}75 - 1{,}50^2 = 0{,}50.$

$E(X^2) = \sum X_i^2\, f(X_i) = 1^2(0{,}625) + 2^2(0{,}250) + 3^2(0{,}125) = 2{,}75.$

6.24. Lo que ocurre con dos ratas es fácil de saber si se considera que el resultado que consiguen entre las dos (Y = «número de bolas») no es más que la suma del número de bolas que consigue cada una por separado. Es decir, $Y = X + X$. Ahora bien, la media de una suma es la suma de las medias (ver, en el Capítulo 4, la nota a pie de página número 2); por tanto: $E(Y) = 2\,E(X) = 2\,(1{,}50) = 3$. Y, cuando dos variables son independientes (como es el caso, pues las ratas recorren independientemente el laberinto), la varianza de una suma es la suma de las varianzas; por tanto, $V(Y) = 2\,V(X) = 2\,(0{,}50) = 1$.

No obstante, también puede obtenerse la distribución muestral de Y y calcular a partir de ella el valor esperado y la varianza de Y:

Y	*Salidas*	*Probabilidades*	$f(Y)$
2	AA	0,625 (0,625)	0,390625
3	AB, BA	0,625 (0,25) + 0,25 (0,625)	0,312500
4	AC, CA, BB	0,625 (0,125) + 0,125 (0,625) + 0,25 (0,25)	0,218750
5	BC, CB	0,25 (0,125) + 0,125 (0,25)	0,062500
6	CC	(0,125) (0,125)	0,015625

b. $E(Y) = \sum Y_i\, f(Y_i) = 2\,(0{,}390625) + 3\,(0{,}312500) + \cdots + 6\,(0{,}015625) = 3.$

$V(Y) = E(Y^2) - [E(Y)]^2 = 10 - 3^2 = 1.$

$$E(X^2) = \sum X_i^2\, f(X_i) = 2^2(0{,}390625) + 3^2(0{,}312500) + \cdots + 6^2(0{,}015625) = 10.$$

7

Introducción a la inferencia estadística I. La estimación de parámetros

Qué es la inferencia estadística

Ya hemos señalado que el análisis de datos es un proceso que se desarrolla en fases: comienza con la selección y recopilación de los datos, continúa con la aplicación de herramientas descriptivas para explorar, organizar y resumir la información contenida en los datos y termina (no necesariamente, pero sí habitualmente) con la aplicación de herramientas inferenciales para llevar a cabo comparaciones y estudiar relaciones. En los capítulos previos se ha estudiado ya lo relativo a la selección de casos (brevemente, pues esta parte es más bien objeto de los diseños de investigación) y a las herramientas disponibles para abordar la fase descriptiva. Éste y los siguientes capítulos se ocupan de la fase inferencial.

La inferencia estadística es un tipo de razonamiento que procede de lo particular a lo general: intenta extraer conclusiones de tipo general a partir de unos pocos datos particulares. Ha llegado el momento de concretar que al hablar de *conclusiones de tipo general* nos estamos refiriendo a conclusiones sobre la forma de una *población* o alguno de sus *parámetros*, y al hablar de *datos particulares* nos estamos refiriendo a una *muestra* de esa población o a alguno de sus *estadísticos*. Y esto, en la práctica, significa, según tendremos ocasión de comprobar repetidamente, *realizar comparaciones* y *estudiar relaciones*.

Estas inferencias (comparaciones, relaciones) pueden realizarse utilizando dos estrategias distintas: la **estimación de parámetros** y el **contraste de hipótesis**. Ambas formas de inferencia son equivalentes en el sentido de que ambas permiten abordar el mismo tipo de problemas y llegar a las mismas conclusiones (podría pensarse en ellas como en las dos caras de una misma moneda), pero la información que ofrecen no es exactamente la misma.

El contraste de hipótesis ha constituido tradicionalmente la esencia de lo que hoy llamamos análisis estadístico; sin embargo, raramente se ha visto libre de críticas (ver Morrison y Henkel, 1970). Estas críticas han alcanzado su máxima expresión en la pasada década de los noventa, la cual ha sido testigo de un agrio debate promovido por una corriente muy crítica con el uso y abuso de esta estrategia (ver, por ejemplo, Chow, 1996; Hagen, 1997; Harlow, Mulaik y Steiger, 1997). Algunos autores han llegado a proponer, incluso, el abandono del

contraste de hipótesis como estrategia de análisis de datos por no considerarlo un método válido para generar conocimiento científico (Cohen, 1990, 1994; Gigerenzer, 1993; Oakes, 1986; Schmidt, 1996). Dejando a un lado estas posturas extremas (difícilmente sostenibles; ver, por ejemplo, Cortina y Dunlap, 1997), las conclusiones de este debate podrían resumirse, quizá, en la recomendación de acompañar todo contraste con su correspondiente estimación.

Lo cierto es que la estimación y el contraste se complementan. A pesar de que son equivalentes en muchos aspectos, la información que ofrecen es algo distinta: mientras que el contraste de hipótesis pone el énfasis en intentar detectar la presencia de un efecto significativo (una diferencia entre grupos, una relación entre variables), la estimación de parámetros pone el énfasis en intentar cuantificar el tamaño de ese efecto (cómo de grande es la diferencia entre los grupos, cómo de intensa es la relación entre las variables).

En este capítulo se explica la lógica de la estimación de parámetros y en el próximo la del contraste de hipótesis. Ambas estrategias son posibles gracias a las distribuciones muestrales (ver capítulo anterior). Tanto el contraste de hipótesis como la estimación de parámetros se basan en la *variabilidad inherente a todo estadístico* (recordemos que un estadístico no es una *constante*, sino una *variable*; en caso necesario, revisar el concepto de *estadístico* en el Capítulo 2). Y son las distribuciones muestrales las que informan sobre esa variabilidad.

Estimación puntual

La **estimación de parámetros** se refiere al *proceso mediante el cual la información muestral es utilizada para inferir valores poblacionales*. Estos valores poblacionales son, lógicamente, parámetros y, es obvio decirlo, la estimación sólo tiene sentido con parámetros desconocidos (pues, si se conoce un parámetro, no es necesario estimarlo). Podemos llevar a cabo dos tipos de estimación: *puntual* y *por intervalos*.

A partir de ahora, para representar un parámetro cualquiera (μ_Y, σ_Y, π_1, etc.), utilizaremos la letra griega θ (theta). Y a los estadísticos utilizados para efectuar estimaciones ($\bar{Y}$, S_Y, P_1, etc.) los llamaremos **estimadores** y los representaremos mediante $\hat{\theta}$.

La estimación puntual constituye la más simple de las inferencias estadísticas que pueden realizarse. Consiste, simplemente, en *asignar un valor muestral concreto al valor poblacional que se desea estimar*. Esto es lo que se hace, por ejemplo, cuando para conocer la edad media de una grupo se toma una muestra aleatoria de ese grupo y se adopta como edad media de todo el grupo la edad media obtenida en la muestra; o cuando, para conocer la proporción de españoles que tienen intención de votar en las próximas elecciones, se toma una muestra aleatoria de españoles y se adopta como proporción poblacional la proporción muestral obtenida.

Ahora bien, el valor muestral concreto asignado en la estimación depende del *método de estimación* adoptado. Uno de los métodos más simples, ideado por Pearson y llamado *método de los momentos*, consiste precisamente en atribuir al parámetro que se desea estimar el valor que toma su correspondiente estadístico. Esto es lo que se hace, por ejemplo, cuando se utiliza la media muestral $\bar{Y}$ para estimar la media poblacional μ_Y, la proporción muestral P_1 para estimar la proporción poblacional π_1, etc. (el Apéndice 7 describe otros métodos de estimación).

El problema de esta estrategia es que, dado un parámetro, siempre es posible definir más de un estadístico diferente para efectuar una estimación del mismo. Puesto que un estadístico es un valor numérico descriptivo de alguna característica muestral, la cantidad de estadísticos que es posible calcular en una muestra cualquiera es prácticamente ilimitada (además de los

que ya conocemos, podríamos definir estadísticos tan pintorescos como "el valor que ocupa el tercer lugar en la muestra", "el logaritmo del inverso del quinto valor muestral", etc.).

Es cierto que existen unos cuantos estadísticos cuya utilidad ha sido repetidamente contrastada. Es cierto también que cualquier parámetro que se desee estimar (μ_Y, σ_Y, π_1, etc.) siempre tiene en la muestra un estadístico paralelo ($\bar{Y}$, S_Y, P_1, etc.). Sin embargo, dada la definición de estadístico (valor numérico que describe una característica muestral), siempre resulta posible, en una muestra cualquiera, definir múltiples estadísticos diferentes. Y no existe una forma *natural* de determinar cuál de todos ellos es el *ideal* para efectuar una estimación correcta. La media poblacional, por ejemplo, podría ser estimada con la media aritmética, o la media recortada, o la mediana, o cualquier otro estadístico de tendencia central de los estudiados en el Capítulo 4. Se hace necesario, por tanto, determinar cuáles son las propiedades que debe reunir un estadístico para poder ser considerado un *buen estimador*.

Propiedades de un buen estimador

Entre las propiedades que debe tener un buen estimador cabe destacar dos[1]: carencia de sesgo y eficiencia. Lo primero que debe exigirse a un estimador es que ofrezca estimaciones *correctas*. No obstante, dado que un estimador (un estadístico) no siempre toma el mismo valor (su valor, ya lo sabemos, depende de la muestra concreta en la que se calcula), no todos los valores que puede tomar coincidirán con el del parámetro estimado. Aun así, de un buen estimador cabe esperar que ofrezca estimaciones correctas al menos en *promedio*. A esta propiedad de ofrecer, en promedio, estimaciones correctas se le llama **carencia de sesgo** y, se dice, por tanto, que un estimador es *insesgado* si su valor esperado (su media) coincide con el parámetro que estima, es decir, si

$$E(\hat{\theta}) = \theta \qquad [7.1]$$

Por ejemplo, los estadísticos $\bar{Y}$, S_Y^2 y P_1 son estimadores insesgados[2] de sus correspondientes parámetros μ_Y, σ_Y^2 y π_1 (de ahí que al estadístico S_Y^2 se le llame varianza *insesgada*). Por el contrario, el estadístico $S_{Y(n)}^2$ es un estimador sesgado de σ_Y^2 (ver, por ejemplo, Pardo y San Martín, 1998, págs. 72-73). Y el *coeficiente de correlación de Pearson* (un estadístico muy

---

[1] Otras propiedades deseables en un estimador son la *consistencia*, la *suficiencia* y la *robustez*. Un estadístico $\hat{\theta}$ es un estimador ***consistente*** del parámetro θ si, para n tendiendo a infinito, se verifica $P(|\hat{\theta} - \theta| < k) \rightarrow 1$ (para una cantidad k infinitamente pequeña). La consistencia garantiza que, a medida que el tamaño muestral va aumentando, también va aumentando la probabilidad de que el valor del estimador coincida exactamente con el parámetro estimado. Los estadísticos *media*, *varianza* (la sesgada y la insesgada) y *proporción* son, todos ellos, estimadores consistentes de sus respectivos parámetros; para constatar esto basta considerar que el tamaño muestral forma parte del denominador de sus respectivas varianzas.

Un estadístico $\hat{\theta}$ es un estimador ***suficiente*** si al estimar el parámetro θ utiliza *toda la información muestral* relacionada con θ. Es decir, si $\hat{\theta}$ es un estimador suficiente de θ, la estimación obtenida no puede mejorarse considerando información muestral no incluida en $\hat{\theta}$. Los estadísticos *media*, *varianza* (la sesgada y la insesgada) y *proporción* son estimadores suficientes de sus respectivos parámetros, pues en todos ellos se utiliza *toda* la información muestral (una simple inspección de sus fórmulas permite comprobar que todos ellos se obtienen a partir de *todos* los elementos muestrales). La moda, sin embargo, se basa en un único valor (el que más se repite); y la mediana se basa sólo en los valores centrales.

La ***robustez*** se refiere al grado en que un estimador se ve afectado cuando no se dan las condiciones óptimas para su aplicación. Decimos que un estadístico es robusto cuando ofrece buenas estimaciones incluso cuando no se dan las condiciones que en teoría deberían darse para ello. Tendremos ocasión de volver sobre esta propiedad.

[2] En efecto, $E(\bar{Y}) = \mu_Y$ (ver Apéndice 6, ecuación [6.10]); $E(S_Y^2) = \sigma_Y^2$ (ver Apéndice 6, ecuación [6.15]); y $E(P_1) = \pi_1$ (ver Capítulo 6, ecuación [6.7]).

útil que estudiaremos más adelante) es otro ejemplo de estimador sesgado, pues su valor esperado sólo coincide con el parámetro que estima cuando éste vale cero.

La segunda propiedad deseable en un estimador es la **eficiencia**. Un estimador es *tanto más eficiente cuanto menor es su varianza*. Supongamos que, para estimar θ, contamos con dos estadísticos distintos $\hat{\theta}_1$ y $\hat{\theta}_2$. Decimos que $\hat{\theta}_1$ es más eficiente que $\hat{\theta}_2$ si se verifica:

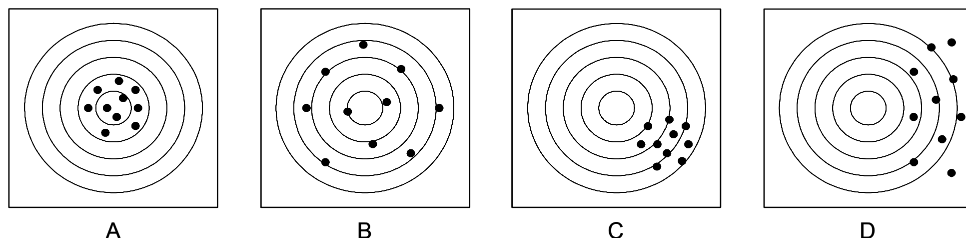
$$\sigma^2_{\hat{\theta}_1} < \sigma^2_{\hat{\theta}_1} \qquad [7.2]$$

Sabemos, por ejemplo, que en una distribución simétrica la media y la mediana toman el mismo valor. Por tanto, para estimar el centro de una distribución simétrica podría utilizarse indistintamente la media ($\bar{Y}$) o la mediana (*Mdn*). Sin embargo, en general, es preferible utilizar $\bar{Y}$ porque, además de ser un estimador insesgado, es más eficiente[3] que *Mdn*. También el parámetro σ^2_Y puede estimarse utilizando dos estadísticos distintos: S^2_n y S^2_{n-1}. Sabemos que sólo S^2_{n-1} es insesgado; pero S^2_n es más eficiente[4] que S^2_{n-1}.

Una mayor eficiencia indica que el estadístico en cuestión varía menos de una muestra a otra, por lo que las estimaciones que pueden hacerse con él son *más precisas* que las que se hacen con un estimador menos eficiente. Lo cierto es que, aunque un estimador *insesgado* ofrece, en promedio, estimaciones correctas, si ese estimador no es *eficiente* (es decir, si su varianza es muy grande) muchas de sus estimaciones estarán muy por encima del verdadero valor del parámetro y otras muchas muy por debajo. Aunque unas y otras se contrarrestarán para ofrecer una estimación promedio correcta, al elegir al azar *una* de esas estimaciones se estará corriendo el riesgo de cometer un error muy grande. De ahí la conveniencia de que un estimador, además de insesgado, sea eficiente.

La Figura 7.1 puede ayudar a comprender el significado de estas dos propiedades. Los cuadros de la figura representan dianas sobre las que se han efectuado 10 disparos. Lógicamente, los disparos se han hecho intentando buscar el centro de la diana. La situación puede extrapolarse fácilmente a la estimación de parámetros si se considera que el centro de la diana representa el parámetro que se desea estimar y que los 10 disparos corresponden a 10 estimaciones efectuadas con un estimador calculado en 10 muestras distintas.

Figura 7.1. Comportamiento de diferentes estimadores: A = insesgado-eficiente; B = insesgado-ineficiente; C = sesgado-eficiente; D = sesgado-ineficiente (adaptado de Wonnacott y Wonnacott, 1990, pág. 242)



[3] Por ejemplo, en el caso concreto de una distribución normal: $\sigma^2_{\bar{Y}} = \sigma^2/n < \sigma^2_{\text{Mdn}} = 1{,}57\sigma^2/n$. Es decir, la varianza del estadístico *media* es menor que la del estadístico *mediana*.

[4] Sabemos que $\sigma^2_{S^2_n} = 2\sigma^4_Y(n-1)/n^2$ y $\sigma^2_{S^2_{n-1}} = 2\sigma^4_Y/(n-1) = 2\sigma^4_Y(n-1)/(n-1)^2$ (ver, por ejemplo, Pardo y San Martín, 1998, págs. 72-73). Ahora bien, como $(n-1)/n^2$ es menor que $(n-1)/(n-1)^2$, puede afirmarse que la *varianza sesgada* es un estimador de σ^2_Y más eficiente que la *varianza insesgada* (lógicamente, esta diferencia en la eficiencia de ambos estimadores va disminuyendo conforme el tamaño muestral va aumentando).

En la diana *A*, los disparos están muy concentrados en torno al blanco; todas las distancias al blanco son bastante cortas y no existe ninguna desviación sistemática del mismo; se trata de un estimador *insesgado* y *eficiente*. En la diana *B*, todos los disparos están muy alejados del blanco; aun no existiendo una desviación sistemática en ninguna dirección, el acierto es bastante escaso; se trata de un estimador *insesgado* pero *poco eficiente*. En la diana *C*, los disparos están concentrados en un punto alejado del blanco; aunque podría decirse que la puntería es bastante alta (los disparos van siempre casi al mismo sitio), existe una desviación sistemática (sesgo) del blanco; se trata de un estimador *eficiente* pero *sesgado*. En la diana *D*, por último, los disparos se encuentran dispersos y alejados del blanco, al igual que en la diana *B*, pero además existe una desviación sistemática (sesgo) del blanco; se trata de un estimador *sesgado* y *poco eficiente*.

Estimación por intervalos

Acabamos de ver que la estimación puntual consiste en atribuir un *valor concreto* al parámetro estimado. Por supuesto, esta forma de hacer inferencia entraña un riesgo evidente, pues el valor que tome un estadístico en una muestra concreta no tiene por qué coincidir exactamente con el valor del parámetro estimado. Debido a la variación muestral, siempre cabe la posibilidad de encontrar cierta discrepancia entre ellos. A esta discrepancia se le llama *error muestral* (*E*) y se define como

$$E = |\hat{\theta} - \theta| \qquad [7.3]$$

En la estimación puntual no hay forma de conocer el valor de *E*; es decir, al utilizar $\hat{\theta}$ como estimador de θ ($\overline{Y}$ como estimador de μ; *P* como estimador de π; etc.) no es posible conocer la magnitud del error que se está cometiendo. Sin embargo, siempre que se efectúa una estimación es importante conocer la *precisión* (y, por tanto, el error) con la que se está trabajando. Esto puede conseguirse recurriendo a la *estimación por intervalos*:

> La **estimación por intervalos** consiste en asignar al parámetro que se desea estimar, no un valor concreto, sino un *rango de valores entre los que se espera que pueda encontrarse el verdadero valor del parámetro con una probabilidad conocida*.

En esta definición hay que destacar la presencia de dos conceptos clave: *rango de valores* y *probabilidad conocida*. Aclaremos estos dos conceptos.

Al rango de valores que se asigna al parámetro se le llama **intervalo de confianza** (*IC*); y a los extremos del intervalo se les llama **límites de confianza**: *límite inferior* (L_i) y *límite superior* (L_s). Para construir un intervalo de confianza para el parámetro θ, al correspondiente estimador puntual $\hat{\theta}$ se le suma y se le resta una cantidad llamada **error máximo** ($E_{\text{máx}}$):

$$IC_\theta = \hat{\theta} \pm E_{\text{máx}} \quad \rightarrow \quad \begin{cases} L_s = \hat{\theta} + E_{\text{máx}} \\ L_i = \hat{\theta} - E_{\text{máx}} \end{cases} \qquad [7.4]$$

El error máximo representa la diferencia máxima que, con una determinada probabilidad, cabe esperar encontrar entre el verdadero valor del parámetro estimado, θ, y el valor concreto del estadístico utilizado para estimarlo, $\hat{\theta}$.

La utilidad de esta estrategia radica justamente en que permite conocer la probabilidad con la que cabe esperar que el intervalo construido incluya el verdadero valor del parámetro estimado. A esa probabilidad se le llama **nivel de confianza** y se representa mediante $1-\alpha$:

$$P(L_i \leq \theta \leq L_s) \quad [7.5]$$

(esta expresión se lee así: la probabilidad de que el parámetro θ se encuentre entre L_i y L_s vale $1-\alpha$). Es claro que para conocer el nivel de confianza es necesario que la cantidad $E_{\text{máx}}$ esté referida a alguna distribución de probabilidad conocida, en concreto, a la distribución muestral del estadístico $\hat{\theta}$ utilizado como estimador. Por tanto, para poder construir intervalos de confianza es necesario utilizar *estimadores con distribución muestral conocida*. Según veremos enseguida, en la estimación por intervalos se comienza fijando el nivel de confianza con el que se desea trabajar (condición de partida) y se continúa calculando la cantidad ($E_{\text{máx}}$) que permite construir el intervalo que satisface la condición impuesta.

Consideremos una población formada por los siguientes elementos $Y_i = \{1, 2, 3, 4, 5\}$. Si extraemos de ella, con reposición, todas las posibles muestras aleatorias de tamaño $n = 2$ y en cada una de ellas calculamos el estadístico $\bar{Y}$, podemos construir la distribución muestral de la media que recoge la Tabla 7.1 (esta población de $N = 5$ elementos ya se ha estudiado en el capítulo anterior, en el apartado *Distribuciones muestrales: un caso concreto*). La tabla incluye las 25 muestras de tamaño $n = 2$ que es posible extraer de la población definida, los distintos valores que puede tomar el estadístico $\bar{Y}$ y la frecuencia relativa asociada a cada uno de ellos. Haciendo los cálculos oportunos se obtiene

$$E(\bar{Y}) = \mu_Y = 3 \qquad \text{y} \qquad \sigma_{\bar{Y}} = \sigma_Y/\sqrt{n} = \sqrt{2}/\sqrt{2} = 1$$

Tabla 7.1. Distribución muestral de la media ($\bar{Y}$) formada a partir de las muestras de tamaño $n = 2$ que es posible extraer con reposición de la población $Y = \{1, 2, 3, 4, 5\}$

Muestras posibles	$\bar{Y}$	$f(\bar{Y})$
(1,1)	1,0	1/25
(1,2) (2,1)	1,5	2/25
(1,3) (2,2) (3,1)	2,0	3/25
(1,4) (2,3) (3,2) (4,1)	2,5	4/25
(1,5) (2,4) (3,3) (4,2) (5,1)	3,0	5/25
(2,5) (3,4) (4,3) (5,2)	3,5	4/25
(3,5) (4,4) (5,3)	4,0	3/25
(4,5) (5,4)	4,5	2/25
(5,5)	5,0	1/25
		1

$\mu_Y = 3$; $\sigma_Y^2 = 2$

Aunque conocemos el valor del parámetro μ_Y y, por tanto, no es necesario estimarlo, vamos a ver qué ocurre al intentar estimarlo. Al seleccionar una muestra de la población definida puede resultar elegida una cualquiera de las 25 posibles. Sabemos que el valor de $\bar{Y}$ va a depender de la muestra concreta elegida. La *estimación por intervalos* consiste en considerar

que el verdadero valor de μ_Y no se alejará del estadístico $\bar{Y}$ en más de una determinada cantidad ($E_{\text{máx}}$). Comencemos suponiendo que esa cantidad es 1 error típico ($\sigma_{\bar{Y}}$), es decir,

$$L_i = \bar{Y} - \sigma_{\bar{Y}} \quad \text{y} \quad L_s = \bar{Y} + \sigma_{\bar{Y}}$$

Al atribuir al parámetro μ_Y un rango de valores comprendidos entre L_i y L_s, el *error máximo* que estamos dispuestos a admitir es de 1 error típico: $E_{\text{máx}} = \sigma_{\bar{Y}}$. Ahora bien, ¿qué garantía tenemos de que nuestra estimación es correcta? Veamos lo que ocurre con cada una de las 25 muestras posibles. Si la muestra que resulta elegida es (1, 1), la media aritmética $\bar{Y}$ valdrá 1, y al construir el intervalo de confianza para efectuar la estimación, como $E_{\text{máx}} = \sigma_{\bar{Y}} = 1$, obtendremos,

$$L_i = 1 - 1 = 0 \quad \text{y} \quad L_s = 1 + 1 = 2$$

lo que nos llevará a inferir que el verdadero valor del parámetro μ_Y se encuentra entre 0 y 2. Como el valor del parámetro es 3, con la muestra (1, 1) estaríamos construyendo un intervalo incorrecto, es decir, estaríamos asignado al parámetro μ_Y un rango de valores entre los que, de hecho, no se encuentra su verdadero valor.

Si en lugar de la muestra (1, 1) resultara elegida la muestra (1, 2) o la muestra (2, 1), el intervalo de confianza se construiría a partir de $\bar{Y}$ = 1,5 y los límites de confianza resultantes serían

$$L_i = 1{,}5 - 1 = 0{,}5 \quad \text{y} \quad L_s = 1{,}5 + 1 = 2{,}5$$

Por tanto, ahora estaríamos afirmando que el verdadero valor de μ_Y se encuentra entre 0,5 y 2,5, lo cual es de nuevo una estimación incorrecta pues el verdadero valor del parámetro μ_Y es 3.

Si resultaran elegidas las muestras (1, 3), (2, 2) o (3, 1), el intervalo de confianza se construiría a partir de $\bar{Y}$ = 2 y los límites de confianza resultantes serían

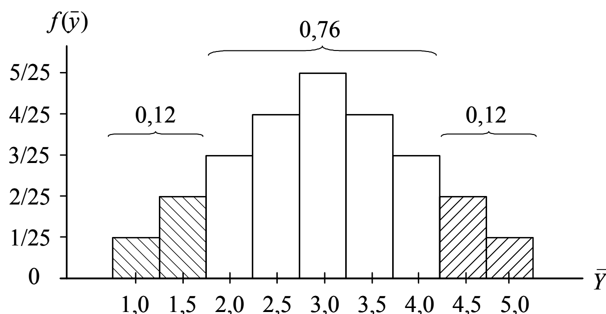
$$L_i = 2 - 1 = 1 \quad \text{y} \quad L_s = 2 + 1 = 3$$

Lo cual significa que con estas tres muestras sí estaríamos construyendo un intervalo correcto, pues el verdadero valor del parámetro (μ_Y = 3) se encuentra entre los valores 1 y 3. Y también construiríamos intervalos correctos con las muestras: (1, 4), (1, 5), (2, 3), (2, 4), (2, 5), (3, 2), (3, 3), (3, 4), (3, 5), (4, 1), (4, 2), (4, 3), (4, 4), (5, 1), (5, 2) y (5, 3), pues con todas ellas el valor de $\bar{Y}$ llevaría a construir intervalos de confianza entre cuyos límites estaría incluido el verdadero valor del parámetro μ_Y. Sin embargo, al igual que con las muestras (1, 1), (1, 2) y (2, 1), también construiríamos intervalos incorrectos (intervalos entre cuyos límites no se encontraría el valor del parámetro μ_Y) con las muestras (4, 5), (5, 4) y (5, 5).

Por tanto, con 19 de las 25 muestras posibles construiríamos intervalos *correctos* y con 6 de las 25 construiríamos intervalos *incorrectos*. Como todas ellas tienen la misma probabilidad de ser elegidas, existe una probabilidad de 19/25 = 0,76 de construir un intervalo que capture el valor de μ_Y; y una probabilidad de 6/25 = 1 − 0,76 = 0,24 de construir un intervalo que no capture el valor de μ_Y.

La Figura 7.2 muestra esto gráficamente: en la *zona no rayada* se encuentran las 19 medias $\bar{Y}$ que llevarán a construir un intervalo correcto; en la *zona rayada* se encuentran las 6 medias $\bar{Y}$ que llevarán a construir un intervalo incorrecto. La anchura o amplitud del intervalo es justamente la parte del eje horizontal correspondiente a la zona no rayada.

Figura 7.2. Distribución muestral de la media formada a partir de las muestras de tamaño n = 2 que es posible extraer de una población de N = 5 elementos



- Llamamos **nivel de confianza** (y lo representamos mediante $1-\alpha$) a la *zona no rayada*: representa la probabilidad de construir un intervalo entre cuyos límites se encuentre el verdadero valor del parámetro.
- Llamamos **nivel de riesgo** o **nivel de significación** (y lo representamos mediante α) a la *zona rayada*: representa la probabilidad de construir un intervalo entre cuyos límites no se encuentre el verdadero valor del parámetro.

Por tanto, al construir un intervalo con $E_{\text{máx}} = \sigma_{\bar{Y}} = 1$, puede afirmarse que el verdadero valor del parámetro μ_Y se encontrará dentro de ese intervalo con un *nivel de confianza* de 0,76 (es decir, con un *nivel de riesgo* de 0,24):

$$P(\bar{Y} - 1 \leq \mu_Y \leq \bar{Y} + 1) = 0{,}76$$

Esto significa que, de todos los intervalos que es posible construir en el escenario descrito, el 76% de ellos incluirá el verdadero valor de μ_Y (el 24% no lo incluirá). Por tanto, trabajando con $E_{\text{máx}} = \sigma_{\bar{Y}}$ puede afirmarse con una confianza de 0,76 (o del 76%) que el intervalo resultante incluye el valor del parámetro μ_Y.

Por supuesto, en lugar de tomar $E_{\text{máx}} = \sigma_{\bar{Y}} = 1$, podría adoptarse cualquier otro valor para $E_{\text{máx}}$. Si, en lugar de un solo error típico, se tomaran 1,5 errores típicos, es decir,

$$E_{\text{máx}} = 1{,}5\,\sigma_{\bar{Y}} = 1{,}5\,(1) = 1{,}5$$

el porcentaje de intervalos que captarían el verdadero valor de μ_Y sería distinto. En concreto, habría 23 muestras de las 25 posibles que llevarían a construir intervalos correctos. Sólo las muestras (1, 1) y (5, 5) (es decir, sólo 2 muestras de las 25 posibles) llevarían a construir intervalos incorrectos (intervalos que no incluirían el valor de μ_Y). Por tanto,

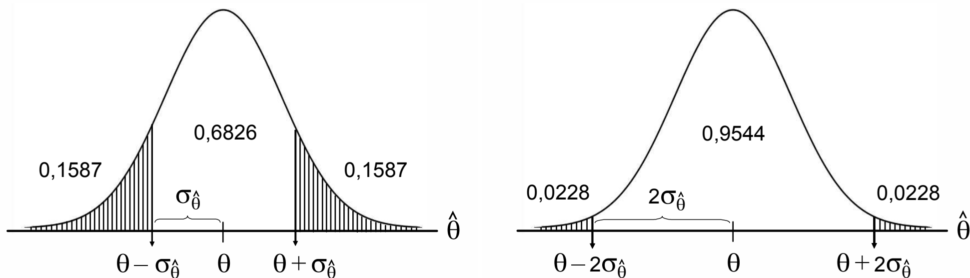
$$P(\bar{Y} - 1{,}5 \leq \mu_Y \leq \bar{Y} + 1{,}5) = 0{,}92$$

Por supuesto, el razonamiento utilizado con $E_{\text{máx}} = \sigma_{\bar{Y}}$ y $E_{\text{máx}} = 1{,}5\,\sigma_{\bar{Y}}$ sigue siendo válido para cualquier otro valor de $E_{\text{máx}}$ ($1{,}96\,\sigma_{\bar{Y}}$, $2\,\sigma_{\bar{Y}}$, $2{,}58\,\sigma_{\bar{Y}}$, etc.). El número de errores típicos que se elija determinará el tamaño de $E_{\text{máx}}$; y el tamaño de $E_{\text{máx}}$ determinará el nivel de confianza del intervalo resultante. Es claro que cuanto mayor sea $E_{\text{máx}}$, mayor será la amplitud del intervalo y mayor también la probabilidad de que el intervalo construido incluya el verdadero valor de θ. Pero debe tenerse en cuenta que, cuanto mayor sea $E_{\text{máx}}$, menor será la precisión

de la estimación, pues se estará atribuyendo al parámetro un rango más amplio de valores. La Figura 7.3 puede ayudar a entender esto con un estadístico distribuido normalmente. La parte del eje horizontal correspondiente a la zona no rayada equivale a la amplitud del intervalo de confianza. Obviamente, si se adopta $E_{\text{máx}} = \sigma_{\hat{\theta}}$ (curva de la izquierda), el intervalo es sensiblemente más estrecho (más preciso) que si se adopta $E_{\text{máx}} = 2\,\sigma_{\hat{\theta}}$ (curva de la derecha), pero ese aumento en la precisión implica una disminución del nivel de confianza.

Estas consideraciones sugieren la necesidad buscar un equilibrio entre dos objetivos contrapuestos: (1) que el intervalo construido sea *lo bastante amplio* como para garantizar que la probabilidad de incluir el parámetro sea alta y, al mismo tiempo, (2) lo *bastante estrecho* como para ofrecer una precisión aceptable. Este equilibrio se ha buscado tradicionalmente en la literatura estadística estableciendo un nivel de confianza de 0,95 y, por tanto, un nivel de riesgo de 0,05. Se consiguen así intervalos de confianza con una precisión aceptable al tiempo que se mantiene un nivel de riesgo razonablemente pequeño. Por supuesto, pueden utilizarse otros niveles de confianza (0,90, 0,99, etc.), pero debe tenerse en cuenta que el nivel de confianza elegido determina el número de errores típicos que es necesario utilizar y, con él, el tamaño de $E_{\text{máx}}$ y la amplitud o precisión del intervalo.

Figura 7.3. Probabilidad asociada a los valores $\pm\sigma_{\hat{\theta}}$ y $\pm 2\sigma_{\hat{\theta}}$ con un estimador $\hat{\theta}$ distribuido normalmente



Dado que las poblaciones con las que se suele trabajar son más grandes que la de este sencillo ejemplo, se comprenderá que no es tarea fácil (ni, muchas veces, posible) encontrar todas las muestras de tamaño n que es posible extraer de ellas. Pero en realidad esto no es un problema porque, para construir intervalos de confianza, todo lo que se necesita es conocer la *distribución muestral* del estadístico utilizado como estimador: la distribución muestral de un estimador informa de la probabilidad asociada a cada uno de sus valores y eso es todo lo que hace falta para seguir la estrategia descrita. Y ya sabemos (ver capítulo anterior) que para conocer la distribución muestral de un estadístico no es necesario seleccionar una sola muestra; existen procedimientos matemáticos que permiten conocer con exactitud el valor esperado, el error típico y la forma de muchas distribuciones muestrales.

Cómo interpretar un intervalo de confianza

Tras calcular un intervalo con un nivel de confianza de 0,95 se suele caer en la tentación de concluir que el parámetro estimado se encuentra entre los límites obtenidos con una *probabilidad* de 0,95. Pero esta afirmación no es del agrado de muchos estadísticos. La razón es que el concepto *probabilidad* debe ir asociado a variables, no a constantes. Y, una vez construido

un intervalo, ya no existe ninguna variable: los límites del intervalo son dos valores concretos, dos constantes, y el parámetro estimado también lo es. Por tanto, al interpretar un intervalo de confianza hay que evitar la palabra *probabilidad*.

Antes de construir el intervalo sí puede hablarse de probabilidad: existe una probabilidad de 0,95 de que la expresión $IC_\theta = \hat{\theta} \pm E_{máx}$ incluya el verdadero valor del parámetro θ. Una vez construido el intervalo ya no tiene sentido hablar de probabilidad: el valor del parámetro se encontrará o no dentro de los límites concretos del intervalo.

Por tanto, un intervalo construido con una confianza de 0,95 puede interpretarse de la siguiente manera: estimamos, con una confianza del 95%, que el verdadero valor del parámetro estimado se encuentra entre los límites del intervalo construido (hemos evitado la palabra *probabilidad*). Y lo que esto significa realmente es que se ha utilizado un procedimiento que permite afirmar que de cada 100 intervalos que se construyan en las mismas condiciones, 95 de ellos incluirán el verdadero valor del parámetro (cinco de ellos no lo harán). Por supuesto, creemos (tenemos una confianza del 95%) que nuestro intervalo es uno de los correctos.

Intervalo de confianza para el parámetro *media*

Sabemos (ver, en el capítulo anterior, el apartado *Distribución muestral del estadístico media*) que, si la variable Y se distribuye normalmente en la población (o si, aun no siendo su distribución normal, el tamaño de la muestra es lo bastante grande), la distribución muestral del estadístico $\bar{Y}$ es normal con $E(\bar{Y}) = \mu_Y$ y $\sigma_{\bar{Y}} = \sigma_Y / \sqrt{n}$. También sabemos que, si el estadístico $\bar{Y}$ se distribuye normalmente, entonces (ver ecuación [6.3]):

$$Z = \frac{\bar{Y} - \mu_Y}{\sigma_{\bar{Y}}} \sim N(0, 1)$$

Ahora bien, en la distribución normal tipificada $N(0, 1)$ se verifica

$$P(Z_{\alpha/2} \leq Z \leq Z_{1-\alpha/2}) = 1 - \alpha \quad \rightarrow \quad P\left(Z_{\alpha/2} \leq \frac{\bar{Y} - \mu_Y}{\sigma_{\bar{Y}}} \leq Z_{1-\alpha/2}\right) = 1 - \alpha \qquad [7.6]$$

A partir de [7.6], unas sencillas transformaciones (ver Pardo y San Martín, 1998, pág. 102) permiten llegar a

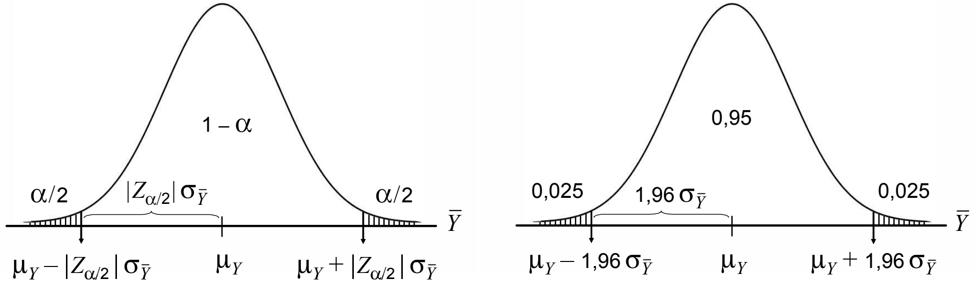
$$P(\underbrace{\bar{Y} - |Z_{\alpha/2}|\,\sigma_{\bar{Y}}}_{L_i} \leq \mu_Y \leq \underbrace{\bar{Y} + |Z_{\alpha/2}|\,\sigma_{\bar{Y}}}_{L_s}) = 1 - \alpha \qquad [7.7]$$

Esto significa que la probabilidad de que el parámetro μ_Y se encuentre entre los valores L_i y L_s, vale $1-\alpha$. Por tanto, haciendo $E_{máx} = |Z_{\alpha/2}|\,\sigma_{\bar{Y}}$ puede afirmarse, con un nivel de confianza de $1-\alpha$, que el valor del parámetro μ_Y no se alejará del estimador $\bar{Y}$ en más de la cantidad $E_{máx}$. En consecuencia, el intervalo de confianza para el parámetro μ_Y puede construirse así:

$$IC_{\mu_Y} = \bar{Y} \pm |Z_{\alpha/2}|\,\sigma_{\bar{Y}} \quad \rightarrow \quad \begin{cases} L_s = \bar{Y} + |Z_{\alpha/2}|\,\sigma_Y/\sqrt{n} \\ L_i = \bar{Y} - |Z_{\alpha/2}|\,\sigma_Y/\sqrt{n} \end{cases} \qquad [7.8]$$

La Figura 7.4 puede ayudar a entender la situación. La curva de la izquierda representa el caso genérico; la de la derecha, el caso concreto en el que $1-\alpha$ vale 0,95. Cualquier media de la zona rayada llevará a construir intervalos que no captarán el valor del parámetro μ_Y, lo que ocurrirá con probabilidad $\alpha = 0{,}025 + 0{,}025 = 0{,}05$. Por el contrario, cualquier media de la zona no rayada llevará a construir intervalos que captarán el valor del parámetro μ_Y, lo que ocurrirá con probabilidad $1-\alpha = 0{,}95$. Es claro que la cantidad $E_{máx}$, viene determinada por el nivel de confianza elegido; con un nivel de confianza de 0,95, $E_{máx} = 1{,}96\,\sigma_{\bar{Y}}$.

Figura 7.4. Curva normal tipificada *N*(0, 1). Valores relevantes en la estimación por intervalos



Si no se conoce el valor de σ_Y, no es posible transformar $\bar{Y}$ en *Z*. Pero sí es posible sustituir σ_Y por su valor muestral para obtener la transformación (ver ecuación [6.4]):

$$T = \frac{\bar{Y} - \mu_Y}{\hat{\sigma}_{\bar{Y}}} \qquad \left(\text{con } \hat{\sigma}_{\bar{Y}} = \frac{S_Y}{\sqrt{n}}\right)$$

La cual ya sabemos que se distribuye según el modelo de probabilidad *t* de *Student* con $n-1$ grados de libertad (ver, en el Capítulo 6, el apartado *Distribución muestral del estadístico media*)[5]. Y, al igual que ocurre en una distribución normal, en una distribución *t* de *Student* se verifica (siguiendo el mismo razonamiento que con la transformación *Z* y la curva normal):

$$P(\underbrace{\bar{Y} - |t_{n-1;\,\alpha/2}|\,\hat{\sigma}_{\bar{Y}}}_{L_i} \leq \mu_Y \leq \underbrace{\bar{Y} + |t_{n-1;\,\alpha/2}|\,\hat{\sigma}_{\bar{Y}}}_{L_s}) = 1 - \alpha \qquad [7.9]$$

[5] En todo momento se está asumiendo población infinita o muestreo aleatorio con reposición en población finita. Recordemos que, en condiciones de muestreo aleatorio sin reposición en población finita, el error típico de la distribución muestral de la media (tanto si se conoce σ_Y como si no) necesita ser corregido:

$$\sigma_{\bar{Y}} = \frac{\sigma_Y}{\sqrt{n}}\sqrt{(N-n)/(N-1)} \qquad (\text{conocida } \sigma_Y)$$
$$\sigma_{\bar{Y}} = \frac{S_Y}{\sqrt{n}}\sqrt{(N-n)/(N-1)} \qquad (\text{desconocida } \sigma_Y) \qquad [7.10]$$

El procedimiento para construir un intervalo de confianza para la media sigue siendo el mismo. Sólo debe tenerse en cuenta que, al trabajar con una población finita de tamaño *N* y muestreo sin reposición, el error típico de la distribución muestral de la media debe corregirse según acabamos de señalar. Por supuesto, a medida que vaya aumentando *N*, el término corrector $(N-n)/(n-1)$ irá tendiendo a 1, lo que significa que muestrear sin reposición una población finita *grande* será equivalente a muestrearla con reposición.

Por tanto, haciendo $E_{\text{máx}} = |t_{n-1;\,\alpha/2}|\,\hat{\sigma}_{\bar{Y}}$, puede afirmarse, con un nivel de confianza de $1-\alpha$, que el valor del parámetro μ_Y no se alejará del estimador $\bar{Y}$ en más de la cantidad $E_{\text{máx}}$. En consecuencia, si se desconoce el valor de σ_Y, el intervalo de confianza para el parámetro μ_Y puede construirse de la siguiente manera:

$$IC_{\mu_Y} = \bar{Y} \pm |t_{n-1;\,\alpha/2}|\,\hat{\sigma}_{\bar{Y}} \quad \rightarrow \quad \begin{cases} L_s = \bar{Y} + |t_{n-1;\,\alpha/2}|\,S_Y/\sqrt{n} \\ L_i = \bar{Y} - |t_{n-1;\,\alpha/2}|\,S_Y/\sqrt{n} \end{cases} \qquad [7.11]$$

Cuando se muestrean poblaciones normales, la distribución t es muy parecida a la normal. Y en condiciones de no-normalidad, los valores de la distribución t se van pareciendo más y más a los de la distribución normal conforme el tamaño muestral va aumentando. Por tanto, con muestras grandes, utilizar Z o T no hace cambiar mucho las cosas. En el Apéndice 5 se estudia con algo de detalle la distribución t y su parecido con la distribución normal tipificada. En el Apéndice 9 se discute el parecido (diferencia) entre las transformaciones Z y T.

Ejemplo. Intervalo de confianza para el parámetro media

Una muestra aleatoria de 100 estudiantes universitarios responde a una prueba de inteligencia espacial (Y) en la que obtiene una media de 80 y una desviación típica insesgada de 10. ¿Entre qué límites cabe esperar que se encuentre la verdadera inteligencia espacial media de los estudiantes universitarios, con un nivel de confianza de 0,95?

Puesto que se desconoce el valor de σ, no es posible utilizar la estrategia propuesta en [7.8] basada en la transformación Z; es necesario recurrir a la estrategia propuesta en [7.11] basada en la transformación T. Además, es necesario asumir que la distribución de la variable Y es normal, lo cual no representa ningún problema porque el tamaño muestral ($n = 100$) es lo bastante grande:

1. $\alpha = 0{,}05$.
2. $|t_{n-1;\,\alpha/2}| = |t_{99;\,0{,}025}| = 1{,}984$.
3. $\hat{\sigma}_{\bar{Y}} S_Y/\sqrt{n} = 10/\sqrt{100} = 1$.
4. $E_{\text{máx}} = |t_{99;\,0{,}025}|\,\hat{\sigma}_{\bar{Y}} = 1{,}984(1) = 1{,}984$.
5. $L_i = 80 - 1{,}984 = 78{,}02$.
   $L_s = 80 + 1{,}984 = 81{,}98$.

Puesto que el tamaño muestral es lo bastante grande, también puede utilizarse la estrategia propuesta en [7.8] basada en la transformación Z:

1. $\alpha = 0{,}05$.
2. $|Z_{\alpha/2}| = |Z_{0{,}025}| = 1{,}96$.
3. $\hat{\sigma}_{\bar{Y}} = S_Y/\sqrt{n} = 10/\sqrt{100} = 1$.
4. $E_{\text{máx}} = |Z_{0{,}025}|\,\hat{\sigma}_{\bar{Y}} = 1{,}96(1) = 1{,}96$.

5. $L_i = 80 - 1{,}96 = 78{,}04$.
$L_s = 80 + 1{,}96 = 81{,}96$.

Puede comprobarse que, efectivamente, siendo *n* grande, el resultado que se obtiene con la distribución normal es muy parecido al que se obtiene con la distribución *t* de *Student*.

Intervalo de confianza para el parámetro *proporción* (IC_π)

Ya sabemos (ver, en el capítulo anterior, el apartado *Distribución muestral del estadístico proporción*) que la distribución muestral del estadístico proporción ($P_{\text{éxito}}$ o P_1) tiende a aproximarse a la distribución normal a medida que *n* va aumentando, con parámetros

$$E(P_1) = \mu_{P_1} = \pi_1$$
$$V(P_1) = \sigma^2_{P_1} = \pi_1(1-\pi_1)/n \quad \rightarrow \quad \sigma_{P_1} = \sqrt{\pi_1(1-\pi_1)/n} \qquad [7.12]$$

Por tanto (ver ecuación [5.8]),

$$Z = \frac{P_1 - \pi_1}{\sigma_{P_1}} \sim N(0,\, 1)$$

Ahora bien, si la transformación Z se distribuye $N(0, 1)$, entonces se verifica

$$P(Z_{\alpha/2} \leq \frac{P_1 - \pi_1}{\sigma_{P_1}} \leq Z_{1-\alpha/2}) = 1 - \alpha \qquad [7.13]$$

A partir de [7.13], unas sencillas transformaciones permiten llegar a

$$P(\underbrace{P_1 - |Z_{\alpha/2}|\sigma_{P_1}}_{L_i} \leq \pi_1 \leq \underbrace{P_1 + |Z_{\alpha/2}|\sigma_{P_1}}_{L_s}) = 1 - \alpha \qquad [7.14]$$

Es decir, la probabilidad de que el parámetro π_1 se encuentre entre L_i y L_s vale $1 - \alpha$. Por tanto, haciendo $E_{\text{máx}} = |Z_{\alpha/2}|\sigma_{P_1}$ puede afirmarse, con un nivel de confianza de $1 - \alpha$, que el valor del parámetro π_1 no se alejará del estimador P_1 en más de la cantidad $E_{\text{máx}}$. En consecuencia, el intervalo de confianza para el parámetro π_1 puede calcularse mediante:

$$IC_{\pi_1} = P_1 \pm |Z_{\alpha/2}|\, \sigma_{P_1} \qquad [7.15]$$

El problema de este intervalo es que, para construirlo, es necesario conocer el error típico de P_1, es decir, σ_{P_1}. Y ocurre que para conocer σ_{P_1} es necesario conocer justamente el parámetro que se desea estimar, es decir, π_1 (ver ecuación [7.12]). No obstante, puesto que P_1 es un estimador consistente del parámetro π_1 (ver la nota a pie de página número 1 de este mismo capítulo), a medida que el tamaño muestral *n* vaya aumentando el valor de P_1 se irá pareciendo

más y más al parámetro π_1 y, consecuentemente, podrá utilizarse en su lugar (ver Pardo y San Martín, 1998, págs. 109-110). En cuyo caso, el intervalo de confianza para el parámetro π_1 podrá construirse mediante:

$$IC_{\pi_1} = P \pm |Z_{\alpha/2}|\,\hat{\sigma}_P \quad \rightarrow \quad \begin{cases} L_s = P + |Z_{\alpha/2}|\sqrt{P(1-P)/n} \\ L_i = P - |Z_{\alpha/2}|\sqrt{P(1-P)/n} \end{cases} \qquad [7.16]$$

Ejemplo. Intervalo de confianza para el parámetro proporción

En una muestra aleatoria de 320 madrileños mayores de 15 años se ha encontrado un 32 % de fumadores. A partir de este dato, ¿entre qué límites cabe esperar que se encuentre, con una confianza del 95%, la verdadera proporción de fumadores en la población de madrileños mayores de 15 años?

Tenemos una variable dicotómica (fumadores-no fumadores) de la que se han realizado $n = 320$ ensayos con proporción de fumadores $P_1 = 0{,}32$. Por tanto,

1. $\alpha = 0{,}05$.
2. $|Z_{\alpha/2}| = |Z_{0,025}| = 1{,}96$.
3. $\hat{\sigma}_{P_1} = \sqrt{P_1(1-P_1)/n} = \sqrt{0{,}32(1-0{,}32)/320} = 0{,}026$.
4. $E_{\text{máx}} = 1{,}96(0{,}026) = 0{,}05$.
5. $L_i = 0{,}32 - 0{,}05 = 0{,}27$.
   $L_s = 0{,}32 - 0{,}05 = 0{,}37$.

Puede afirmarse, con una confianza del 95%, que la proporción de fumadores en la población de madrileños mayores de 15 años se encuentra entre 0,27 y 0,37.

¿Qué amplitud tendría el intervalo con una muestra de mayor tamaño? Ya hemos señalado que, al aumentar el tamaño muestral, disminuye la amplitud del intervalo. Con $n = 500$, por ejemplo, se obtienen unos límites de 0,28 y 0,36.

Apéndice 7

Precisión de la estimación y tamaño de la muestra

Un intervalo de confianza suele ser tanto más útil e informativo cuanto menor es su amplitud. Ahora bien, la amplitud de un intervalo depende de dos factores: el nivel de confianza utilizado y el error típico del estimador. Si disminuye el nivel de confianza, también lo hace la amplitud del intervalo, pero a costa de incrementar el riesgo, lo cual no parece una solución razonable. La reducción de la amplitud del intervalo hay que intentar conseguirla sin alterar el nivel de confianza; y eso pasa, necesariamente,

por la reducción del error típico del estimador: cualquier acción que pueda llevarse a cabo para reducir el error típico tendrá como consecuencia una reducción de la amplitud del intervalo.

En el caso de la media, su error típico depende tanto de la varianza de la población como del tamaño de la muestra, pues $\sigma_{\bar{Y}} = \sigma_Y/\sqrt{n}$. Por lo que se refiere a la varianza poblacional, aunque es cierto que no es posible eliminarla por completo porque las fuentes de error en un estudio empírico son muchas y de muy diversa índole, una cuidadosa elaboración del diseño de investigación puede contribuir de forma eficaz a conseguir una importante reducción de la misma. Por lo que se refiere al tamaño de la muestra, es claro que un incremento del mismo tiene como consecuencia directa una disminución del error típico $\sigma_{\bar{Y}}$. Lo cual implica que, modificando el tamaño de la muestra, *es posible controlar el grado de precisión del intervalo*.

Veamos qué puede hacerse con el tamaño de la muestra para conseguir disminuir el error típico y obtener, como consecuencia de ello, una mayor precisión en la estimación. De acuerdo con el teorema de Tchebychev (ver, por ejemplo, Amón, 1984, págs. 130-131):

$$P\left(\left|\frac{\hat{\theta}-\theta}{\sigma_{\hat{\theta}}}\right| > |k|\right) < \frac{1}{k^2} \qquad \rightarrow \qquad P(|\hat{\theta}-\theta| > |k|\sigma_{\hat{\theta}}) < \frac{1}{k^2} \qquad [7.17]$$

Conocida la distribución muestral del estimador $\hat{\theta}$ y siendo k un valor estandarizado de la misma:

$$P(|\hat{\theta}-\theta| > |k_{\alpha/2}|\sigma_{\hat{\theta}}) < \frac{1}{k^2_{\alpha/2}} = \alpha$$

de donde, para un nivel de confianza $1-\alpha$ dado, tendremos

$$E = |k_{\alpha/2}|\,\sigma_{\hat{\theta}} \qquad \rightarrow \qquad E^2 = k^2_{\alpha/2}\,\sigma^2_{\hat{\theta}} \qquad \rightarrow \qquad \sigma^2_{\hat{\theta}} = \frac{E^2}{k^2_{\alpha/2}} \qquad [7.18]$$

A partir de estas ecuaciones es posible determinar cuál ha de ser el tamaño de la muestra para alcanzar una determinada precisión. Consideraremos el caso de la media y el de la proporción.

El caso de la media

Sabemos que $\sigma^2_{\bar{Y}} = \sigma^2_Y/n$. Haciendo $k = z$, se obtiene

$$\frac{\sigma^2_Y}{n} = \frac{E^2}{Z^2_{\alpha/2}} \qquad \rightarrow \qquad n = \sigma^2_Y\,\frac{Z^2_{\alpha/2}}{E^2} \qquad [7.19]$$

que, para un nivel de riesgo dado, indica el tamaño muestral necesario (n) para obtener una precisión concreta (E). Si se desconoce σ^2, sabemos que la tipificación del estadístico *media* no sigue la distribución normal, sino la distribución t de *Student*. En tal caso, haciendo $k = t$, se obtiene

$$\frac{S^2_Y}{n} = \frac{E^2}{t^2_{n-1;\,\alpha/2}} \qquad \rightarrow \qquad n = S^2_Y\,\frac{t^2_{n-1;\,\alpha/2}}{E^2} \qquad [7.20]$$

Recordemos el ejemplo utilizado anteriormente en el apartado sobre el intervalo de confianza para el parámetro μ_Y. Con una muestra de 100 estudiantes universitarios, una media muestral de 80, una desviación típica insesgada de 10 y un nivel de confianza de 0,95, se construyó un intervalo de confianza con límites 78,02 y 81,98, es decir un intervalo con una amplitud de 81,98 − 78,02 = 3,96 puntos. Sin cambiar el nivel de confianza (0,95), ¿qué tamaño muestral sería necesario para que el intervalo construido tuviera una amplitud de 2 puntos (es decir, un error máximo de 1 punto)?

Con la muestra de 100 sujetos, el intervalo tiene una amplitud de 3,96 puntos. El tamaño muestral necesario para conseguir una amplitud de 2 puntos es

$$n = S_Y^2 \frac{t_{n-1;\,\alpha/2}^2}{E^2} = 10^2 \frac{t_{99;\,0,025}^2}{1^2} = 100 \frac{-1,984^2}{1} = 393,63 \quad \text{(es decir, 394 sujetos)}$$

El caso de la proporción

Recordemos que, con tamaños muestrales grandes, el error típico de la proporción puede estimarse mediante

$$\hat{\sigma}_{P_1} = \sqrt{P_1\,(1 - P_1)/n}$$

En consecuencia:

$$\frac{P_1\,(1-P_1)}{n} = \frac{E^2}{Z_{\alpha/2}^2} \quad \rightarrow \quad n = P_1\,(1-P_1)\,\frac{Z_{\alpha/2}^2}{E^2} \qquad [7.21]$$

Recordemos el ejemplo utilizado en el aparado sobre el intervalo de confianza para el parámetro π. Con una muestra de 320 madrileños mayores de 15 años, un 32% de fumadores y un nivel de confianza de 0,95, se construyó un intervalo con límites 0,27 y 0,37, es decir, con una amplitud de 0,10 puntos. Sin cambiar el nivel de confianza (0,95), ¿qué tamaño muestral sería necesario utilizar para que el intervalo construido tuviera una amplitud de 0,05 puntos? Aplicando la ecuación [7.21] se obtiene:

$$n = P_1(1-P_1)\,\frac{Z_{\alpha/2}^2}{E^2} = 0,32\,(1-0,32)\,\frac{1,96^2}{0,05^2} = 334,4 \quad \text{(es decir, 335 sujetos)}$$

Estimación por máxima verosimilitud

Ya hemos hablado de las propiedades que debe tener un buen estimador. Para encontrar estimadores que posean todas o algunas de esas propiedades existen diferentes métodos de estimación que, aunque sólo sea superficialmente, conviene conocer.

Uno de estos métodos, propuesto por Fisher, se conoce con el nombre de *máxima verosimilitud*. Consiste en seleccionar como estimador de un parámetro el valor capaz de maximizar la verosimilitud del resultado muestral concreto obtenido, entendiendo por *verosimilitud* la probabilidad de, dados uno o más parámetros concretos, obtener el resultado muestral de hecho obtenido.

Consideremos una variable aleatoria X con distribución de probabilidad poblacional conocida (el método de máxima verosimilitud exige conocer la forma de la distribución de probabilidad con la que se va a trabajar) y supongamos que de esa distribución de probabilidad, aunque conocemos la forma, desconocemos el parámetro θ (o los parámetros $\theta_1, \theta_2, ..., \theta_k$). Extraigamos una muestra aleatoria de tamaño n y representemos por $(X_1, X_2, ..., X_n)$ el resultado muestral concreto obtenido.

Llamamos *función de verosimilitud* a la función de probabilidad (o de densidad de probabilidad) conjunta asociada al resultado muestral concreto $(X_1, X_2, ..., X_n)$, dado el parámetro θ:

$$V(X_1, X_2, ..., X_n; \theta) \qquad [7.22]$$

Para cada posible valor de θ, la función de probabilidad del resultado muestral $(X_1, X_2, ..., X_n)$ tendrá un valor, probablemente, distinto. Pues bien, el método de estimación de máxima verosimilitud consiste

justamente en encontrar el valor de θ que hace máxima la probabilidad conjunta de obtener un resultado muestral como el obtenido. Se trata, por tanto, de maximizar V.

Este método de estimación puede ser fácilmente ilustrado utilizando la distribución binomial. Supongamos que, de una determinada población, extraemos una muestra aleatoria de tamaño $n = 20$ y que cada sujeto es clasificado como *hombre* o como *mujer*. Llamemos π_H a la *proporción de hombres en la población*. La variable aleatoria n_H = «número de hombres en la muestra» será una variable distribuida binomialmente con parámetros n y π_H. Aunque conocemos el valor de n, desconocemos el valor de π_H. ¿Cómo estimarlo? Supongamos que, en la muestra elegida, la variable n_H toma el valor 6. ¿Cuál es el valor de π_H que hace más probable el resultado muestral $n_H = 6$? La respuesta a esta pregunta es la estimación de máxima verosimilitud para el parámetro π_H.

Puesto que la variable n_H se distribuye binomialmente, la probabilidad de obtener $n_H = 6$ con los posibles diferentes valores de π_H, puede calcularse de la siguiente manera:

$$P(n_H) = \binom{n}{n_H} \pi_H^{n_H} (1-\pi_H)^{n-n_H} \qquad [7.23]$$

Por supuesto, también podemos utilizar la tabla de la distribución binomial (ver Tabla B del apéndice final). De una u otra forma obtendremos, para $\pi_H = 0{,}10$:

$$P(n_H = 6 \mid \pi_H = 0{,}10) = \binom{20}{6} (0{,}10)^6 (0{,}90)^{20-6} = 0{,}0089.$$

Para $\pi_H = 0{,}20$,

$$P(n_H = 6 \mid \pi_H = 0{,}20) = \binom{20}{6} (0{,}20)^6 (0{,}80)^{20-6} = 0{,}1091.$$

Para $\pi_H = 0{,}30$,

$$P(n_H = 6 \mid \pi_H = 0{,}30) = \binom{20}{6} (0{,}30)^6 (0{,}70)^{20-6} = 0{,}1916.$$

Podemos seguir calculando, para cada posible valor de π_H, la probabilidad de obtener el resultado muestral concreto $n_H = 6$. Pero a partir de $\pi_H = 0{,}30$ esas probabilidades comienzan a disminuir (puede comprobarse fácilmente). De modo que el principio de máxima verosimilitud nos llevará a concluir que el parámetro $\pi_H = 0{,}30$ es el que hace más probable el resultado muestral $n_H = 6$. Por tanto, decidiremos utilizar $\hat{\pi}_H = 0{,}30$ como estimación *maximoverosímil* del parámetro π_H = «proporción de hombres en la población ».

Este sencillo ejemplo sirve para formarnos una idea de cómo funciona el método de estimación de máxima verosimilitud. Pero para conocer cuál es el valor del parámetro que maximiza la probabilidad de un resultado muestral concreto no necesitamos calcular una a una todas las probabilidades de ese resultado muestral bajo todos los posibles valores asumibles por el parámetro en cuestión. Podemos maximizar V utilizando procedimientos matemáticos mucho más directos (ver, por ejemplo, Ríos, 1985, págs. 328-330; o Amón, 1984, págs. 249-254).

Sin embargo, no pretendemos que el lector conozca la forma concreta de obtener una estimación por el método de máxima verosimilitud. Lo que nos interesa destacar aquí es el importante punto de vista general sobre el que descansa el principio o método de máxima verosimilitud. Este punto de vista se refiere a que *las características poblacionales verdaderas deberán ser aquellas que hagan* ***probables*** *nuestros resultados muestrales*. Si una situación teórica convierte en improbables los resultados empíricos obtenidos, deberemos dudar de ella. La razón es sencilla de entender: si una situación teórica hace improbable la aparición de un resultado empírico concreto y, sin embargo, ese resultado empírico se

produce, habrá que pensar que la situación teórica planteada no puede ser verdadera. Las afirmaciones teóricas son *creíbles* en la medida en que los datos empíricos se muestran compatibles con ellas. Por supuesto, los datos de un único experimento nunca deben considerarse definitivos a la hora de confirmar o no una teoría; se requieren varias réplicas, variaciones en el diseño, diferentes tipos de mediciones, etc., y, aun así, la confirmación de una teoría difícilmente se convierte en definitiva; sin embargo, el punto de vista implícito en el principio de máxima verosimilitud siempre está presente en los diferentes procedimientos de análisis de datos y, consecuentemente, en la propia metodología científica.

Estimación por mínimos cuadrados

Otro importante método de estimación (muy útil en ciertos casos; ver capítulo 9) consiste en utilizar como estimación de un parámetro aquel valor que hace mínimas las distancias al cuadrado entre ese valor estimado y los resultados muestrales observados. Este método no requiere conocer la forma de la distribución de probabilidad con la que se está trabajando (como ocurría con el método de máxima verosimilitud) pero no es válido para obtener estimaciones con todo tipo de parámetros.

Consideremos el caso de la media. Extraigamos de una población cualquiera una muestra aleatoria de tamaño n. Llamemos $(Y_1, Y_2, ..., Y_n)$ al resultado muestral concreto obtenido. Siguiendo el método de mínimos cuadrados utilizaremos como estimador de μ_Y el valor $\hat{\mu}_Y$ que haga mínima la expresión

$$\sum_i^n (Y_i - \hat{\mu}_Y)^2 \qquad [7.24]$$

Es decir, utilizaremos como estimador de μ_Y el valor que consiga hacer mínimas las distancias al cuadrado respecto a los n elementos del resultado muestral obtenido. Sumando y restando $\bar{Y}$ en [7.24], agrupando y desarrollando, obtenemos

$$\begin{aligned}\sum_i^n (Y_i - \hat{\mu}_Y)^2 &= \sum_i^n (Y_i - \bar{Y} + \bar{Y} - \hat{\mu}_Y)^2 = \sum_i^n \left[(Y_i - \bar{Y}) + (\bar{Y} - \hat{\mu}_Y)\right]^2 = \\ &= \sum_i^n \left[(Y_i - \bar{Y})^2 + (\bar{Y} - \hat{\mu}_Y)^2 + 2(Y_i - \bar{Y})(\bar{Y} - \hat{\mu}_Y)\right] = \\ &= \sum_i^n (Y_i - \bar{Y})^2 + \sum_i^n (\bar{Y} - \hat{\mu}_Y)^2 + \sum_i^n 2(Y_i - \bar{Y})(\bar{Y} - \hat{\mu}_Y)\end{aligned} \qquad [7.25]$$

Teniendo en cuenta que

$$\sum_i^n 2(Y_i - \bar{Y})(\bar{Y} - \hat{\mu}_Y) = 2(\bar{Y} - \hat{\mu}_Y)\sum_i^n (Y_i - \bar{Y}) = 0, \qquad [7.26]$$

la expresión [7.25] se reduce a

$$\sum_i^n (Y_i - \hat{\mu}_Y)^2 = \sum_i^n (Y_i - \bar{Y})^2 + \sum_i^n (\bar{Y} - \hat{\mu}_Y)^2 \qquad [7.27]$$

Ahora bien, el término

$$\sum_i^n (Y_i - \bar{Y})^2 \qquad [7.28]$$

no es más que numerador de la ya conocida fórmula de la varianza:

$$S_Y^2 = \frac{\sum_i^n (Y_i - \bar{Y})^2}{n - 1} \quad \rightarrow \quad \sum_i^n (Y_i - \bar{Y})^2 = (n - 1)\, S_Y^2 \qquad [7.29]$$

Por tanto, sustituyendo en [7.27] y teniendo en cuenta que tanto $\bar{Y}$ como $\hat{\mu}_Y$ son términos constantes y que el sumatorio de una constante es *n* veces esa constante, se llega a

$$\sum_{i}^{n} (Y_i - \hat{\mu}_Y)^2 = (n-1)\, S_n^2 + (n-1)(\bar{Y} - \hat{\mu}_Y)^2 \qquad [7.30]$$

Puesto que ninguno de los tres términos en [7.30] puede ser negativo (se trata del tamaño muestral y de cantidades elevadas al cuadrado), la suma de las desviaciones $(Y_i - \hat{\mu}_Y)^2$ siempre será menor o igual que $(n-1)\, S_Y^2$. Y esa suma será mínima cuando $(n-1)(\bar{Y} - \hat{\mu}_Y)^2$ valga cero, lo cual sólo es posible cuando $\bar{Y} = \hat{\mu}_Y$. Esto significa que la media muestral $\bar{Y}$ es el *estimador minimocuadrático* de la media poblacional μ_Y.

Ejercicios

7.1. Las puntuaciones *Y* que se obtienen en la *Escala de Inteligencia para Adultos de Wechsler* (WAIS) se distribuyen normalmente con $\mu_Y = 100$ y $\sigma_Y = 15$. Un psicólogo ha construido una nueva prueba de inteligencia y desea saber si la media estandarizada que se obtiene con ella se parece o no a la que proporciona el WAIS. Para ello, selecciona una muestra aleatoria de 100 sujetos y, tras pasarles la prueba, obtiene una media de 104. El psicólogo desea saber si el intervalo de confianza que se obtiene a partir de la nueva media incluye o no entre sus límites el valor original 100. ¿A qué conclusión llegará el psicólogo, con un nivel de significación de 0,05?

7.2. ¿A qué conclusión habría llegado el psicólogo del ejercicio anterior si, con el mismo nivel de confianza, en lugar de seleccionar una muestra de 100 sujetos, hubiera seleccionado una muestra de 30 sujetos?

7.3. Al seleccionar una muestra aleatoria de 40 pacientes de un hospital se ha encontrado que 8 de ellos presentan síntomas depresivos. ¿Entre qué límites cabe esperar, con un nivel de confianza de 0,95, que se encuentre la proporción de pacientes con síntomas depresivos en la población de pacientes de ese hospital?

7.4. Una muestra aleatoria de 100 sujetos ha obtenido en una escala de satisfacción con la atención recibida en el servicio de urgencias de un determinado hospital una media de 10 y una desviación típica de 4. Al calcular los límites de confianza para la media poblacional se ha obtenido una puntuación de 9 para el límite inferior y de 11 para el superior. ¿A qué nivel de confianza se han calculado estos límites?

7.5. De los 150 estudiantes aleatoriamente seleccionados entre los matriculados en una determinada facultad, el 43 % mostró actitudes calificables de *hostiles* (*H*) hacia la asignatura *Análisis de datos*. Con un nivel de confianza de 0,95, ¿entre qué límites estimaremos que se encuentra la verdadera proporción de estudiantes de esa facultad con actitudes hostiles hacia la mencionada asignatura?

7.6. ¿Qué tamaño debería tener la muestra del ejercicio anterior para que, al estimar la verdadera proporción de estudiantes con actitudes hostiles, con el mismo nivel de confianza, el error máximo del intervalo de confianza obtenido sea justamente la mitad del obtenido con la muestra de 150 estudiantes?

7.7. A partir de una muestra aleatoria de 100 estudiantes universitarios hemos obtenido una media de 50 y una varianza de 225 en una prueba de razonamiento abstracto. Queremos estimar la media de la población sirviéndonos de un intervalo de confianza cuyo error máximo sea de 3 puntos. ¿Qué nivel de confianza debemos utilizar?

7.8. En un experimento sobre agudeza visual se ha presentado a un sujeto 50 pares de estímulos luminosos para comprobar si era capaz de percibir la diferencia en intensidad entre los dos estímulos de cada par. El sujeto debía pulsar un botón rojo cuando creía que la intensidad luminosa de los estímulos era distinta y un botón verde cuando creía que la intensidad luminosa de los estímulos era la misma. ¿Con qué número de aciertos (A) podremos pensar que el sujeto no ha estado pulsando los botones al azar? ($\alpha = 0{,}05$).

7.9. Un investigador desea construir una escala con pares de estímulos visuales para utilizarla en experimentos sobre agudeza visual. Desea que la dificultad de discriminación de cada par de estímulos sea de nivel medio. Elabora 25 pares de estímulos y los pasa a 100 sujetos con la intención de seleccionar para su escala aquellos pares que sean percibidos como diferentes (D) por la mitad de los sujetos. ¿Cuál será el número mínimo y máximo de sujetos que tendrán que percibir los estímulos de cada par como *diferentes* para que se cumpla el criterio de dificultad media establecido por el investigador? ($\alpha = 0{,}05$).

7.10. Una prueba de aptitud espacial esta formada por 15 preguntas dicotómicas (dos alternativas de respuesta). ¿Qué número mínimo y máximo de aciertos (A) cabe esperar que tenga un sujeto que responde al azar? ($\alpha = 0{,}05$).

7.11. Algunos trabajos señalan que, en la Comunidad de Madrid, los trastornos depresivos afectan al 32% de las personas en paro. Un psicólogo social sospecha que esa cifra está desfasada y decide obtener nueva información. Selecciona una muestra aleatoria de 300 personas en paro y encuentra que 63 de ellos muestran trastornos depresivos. Utilizando $\alpha = 0{,}05$, ¿a qué conclusión llegará el psicólogo sobre la afirmación inicial de que el 32% de las personas en paro padecen trastornos depresivos?

7.12. Sabemos que la variable Y se distribuye $N(\mu_Y, \sigma_Y)$. Utilizando un nivel de confianza de 0,05 y una muestra aleatoria de tamaño n se ha obtenido el siguiente intervalo de confianza para la media: $L_i = 80$, $L_s = 90$. Esto supuesto (señalar la/s alternativa/s correcta/s):

a. Vale 0,95 la probabilidad de que la media de cualquier muestra aleatoria procedente de esa población caiga dentro del intervalo (80, 90).

b. Vale 0,05 la probabilidad de que el parámetro μ_Y caiga dentro del intervalo (80, 90) con tal de que la muestra sea aleatoria.

c. Extrayendo un número indefinido de muestras aleatorias de tamaño n de esa población, el 95% de los intervalos construidos contendrá el parámetro μ_Y.

d. Extrayendo un número indefinido de muestras aleatorias de tamaño *n* de esa población, vale 0,05 la probabilidad de que uno de ellos no contenga el parámetro μ_Y.

e. Vale 0,95 la probabilidad de que, al seleccionar aleatoriamente una puntuación de esa población, ésta se encuentre dentro del intervalo (80, 90).

7.13. La anchura de un intervalo de confianza (señalar la/s alternativa/s correcta/s):

a. Es independiente del tamaño de la muestra.
b. Aumenta cuando aumenta el tamaño de la muestra.
c. Es mayor cuanto más homogénea es la población.
d. Aumenta cuando aumenta el nivel de confianza $1-\alpha$.
e. Disminuye cuando aumenta el nivel de confianza $1-\alpha$.
f. Disminuye cuando disminuye el error típico del estimador.

7.14. Un estudiante lee en un informe que los límites del intervalo de confianza al 95% para el cociente intelectual medio de los universitarios españoles se encuentra entre 98 y 106. Al preguntar a un compañero por el significado de ese intervalo, éste le responde que "el 95% de los universitarios españoles tiene un cociente intelectual comprendido entre 98 y 106". ¿Ha hecho el estudiante una interpretación correcta del intervalo de confianza? ¿Por qué?

Soluciones

7.1. Intervalo de confianza para el parámetro *media* con $\sigma_Y = 15$.

- $\alpha = 0{,}05$.
- $|Z_{\alpha/2}| = |Z_{0,025}| = 1{,}96$.
- $\sigma_{\bar{Y}} = \sigma_Y/\sqrt{n} = 15/\sqrt{100} = 1{,}5$.
- $E_{\text{máx}} = |Z_{0,025}|\,\hat{\sigma}_{\bar{Y}} = 1{,}96(1{,}5) = 2{,}94$.
- $L_i = 104 - 2{,}94 = 101{,}06$.
  $L_s = 104 + 2{,}94 = 106{,}94$.

El intervalo de confianza no incluye el valor 100. No parece, por tanto, que la media de la nueva escala sea igual que la del WAIS.

7.2. Intervalo de confianza para el parámetro *media* con $\sigma = 15$.

- $\alpha = 0{,}05$.
- $|Z_{\alpha/2}| = |Z_{0,025}| = 1{,}96$.
- $\sigma_{\bar{Y}} = \sigma_Y/\sqrt{n} = 15/\sqrt{30} = 2{,}74$.
- $E_{\text{máx}} = |Z_{0,025}|\,\hat{\sigma}_{\bar{Y}} = 1{,}96(2{,}74) = 5{,}37$.
- $L_i = 104 - 5{,}37 = 98{,}63$.
  $L_s = 104 + 5{,}37 = 109{,}37$.

Ahora, el intervalo de confianza incluye el valor 100. El resultado obtenido es compatible con la suposición de que la media de la nueva escala es igual que la del WAIS.

7.3. Intervalo de confianza para el parámetro *proporción*, con $n = 40$ y $P_D = 8/40 = 0{,}20$.

- $\alpha = 0{,}05$.
- $|Z_{\alpha/2}| = |Z_{0{,}025}| = 1{,}96$.
- $\hat{\sigma}_{P_D} = \sqrt{P_D(1-P_D)/n} = \sqrt{0{,}20(1-0{,}20)/40} = 0{,}06$.
- $E_{\text{máx}} = 1{,}96(0{,}06) = 0{,}12$.
- $L_i = 0{,}20 - 0{,}12 = 0{,}08$.

  $L_s = 0{,}20 + 0{,}12 = 0{,}32$.

7.4. Intervalo de confianza para el parámetro *media* con σ desconocida y $n = 100$.
Aunque se desconoce el valor de σ, la muestra es lo bastante grande como para poder utilizar la distribución normal sin problemas.

- $E_{\text{máx}} = (11-9)/2 = 1$.
- $\hat{\sigma}_{\bar{Y}} = S_Y/\sqrt{n} = 4/\sqrt{100} = 0{,}40$.
- $E_{\text{máx}} = |Z_{\alpha/2}|\,\hat{\sigma}_{\bar{Y}} \quad \rightarrow \quad 1 = |Z_{\alpha/2}|\,0{,}40 \quad \rightarrow \quad |Z_{\alpha/2}| = 1/0{,}40 = 2{,}50$.
- $P(Z \geq 2{,}50) = \alpha/2 = 0{,}0062 \quad \rightarrow \quad \alpha = 2(0{,}0062) = 0{,}0124$.
- $1-\alpha = 1-(0{,}0124) = 0{,}9876$.

7.5. Intervalo de confianza para el parámetro *proporción*, con $n = 150$ y $P_H = 0{,}43$.

- $\alpha = 0{,}05$.
- $|Z_{\alpha/2}| = |Z_{0{,}025}| = 1{,}96$.
- $\hat{\sigma}_{P_H} = \sqrt{P_H(1-P_H)/n} = \sqrt{0{,}43(1-0{,}43)/150} = 0{,}04$.
- $E_{\text{máx}} = 1{,}96(0{,}04) = 0{,}08$.
- $L_i = 0{,}43 - 0{,}08 = 0{,}35$.

  $L_s = 0{,}43 + 0{,}08 = 0{,}51$.

7.6. Intervalo de confianza para el parámetro *proporción* con n desconocido y $P_H = 0{,}43$.

- $E_{\text{máx}} = 0{,}08/2 = 0{,}04$.
- $\hat{\sigma}_{P_H} = \sqrt{P_H(1-P_H)/n} = \sqrt{0{,}43(1-0{,}43)/n} = \sqrt{0{,}2451/n}$.
- $E_{\text{máx}} = |Z_{1-\alpha/2}|\,\hat{\sigma}_{\bar{Y}} \quad \rightarrow \quad 0{,}04 = 1{,}96\sqrt{0{,}2451/n}$.
- $n = 1{,}96^2\,(0{,}2451)/0{,}04^2 \approx 588$.

(Puede ahorrarse algo de trabajo aplicando la ecuación [7.21]).

7.7. Intervalo de confianza para el parámetro *media* con σ desconocida y $n = 100$.
Aunque se desconoce el valor de σ, la muestra es lo bastante grande como para poder utilizar la distribución normal sin problemas.

- $E_{\text{máx}} = 3$.
- $\hat{\sigma}_{\bar{Y}} = S_Y/\sqrt{n} = 15/\sqrt{100} = 1{,}5$.
- $E_{\text{máx}} = |Z_{\alpha/2}|\,\hat{\sigma}_{\bar{Y}} \quad \rightarrow \quad 3 = |Z_{\alpha/2}|\,1{,}5 \quad \rightarrow \quad |Z_{1-\alpha/2}| = 3/1{,}5 = 2{,}0$.
- $P(Z \geq 2{,}0) = \alpha/2 = 0{,}0228 \quad \rightarrow \quad \alpha = 2(0{,}0228) = 0{,}0456$.
- $1-\alpha = 1-(0{,}0456) = 0{,}9544$.

7.8. Intervalo de confianza para el parámetro *proporción* con $n = 50$ y $P_A = 0{,}50$.

- $\alpha = 0{,}05$.
- $|Z_{\alpha/2}| = |Z_{0{,}025}| = 1{,}96$.
- $\sigma_{P_A} = \sqrt{P_A(1-P_A)/n} = \sqrt{0{,}50(1-0{,}50)/50} = 0{,}071$.
- $E_{\text{máx}} = 1{,}96(0{,}071) = 0{,}14$.
- $L_i = 0{,}50 - 0{,}14 = 0{,}36$.

  $L_s = 0{,}50 + 0{,}14 = 0{,}64$.

Pensaremos que el sujeto no está respondiendo al azar cuando tenga menos de un 36 % de aciertos o más de un 64 %. Es decir, con 18 aciertos o menos y con 32 aciertos o más.

7.9. Intervalo de confianza para el parámetro *proporción* con $n = 100$ y $P_D = 0{,}50$.

- $\alpha = 0{,}05$.
- $|Z_{\alpha/2}| = |Z_{0{,}025}| = 1{,}96$.
- $\sigma_{P_D} = \sqrt{P_D(1-P_D)/n} = \sqrt{0{,}50(1-0{,}50)/100} = 0{,}05$.
- $E_{\text{máx}} = 1{,}96(0{,}05) = 0{,}098$.
- $L_i = 0{,}50 - 0{,}098 = 0{,}402$.

  $L_s = 0{,}50 + 0{,}098 = 0{,}598$.

Según el criterio establecido por el investigador, los estímulos dejarán de ser de dificultad media cuando los acierten menos del 40,2 % o más del 59,8 % de los sujetos. Por tanto, elegirá para su escala los pares de estímulos que sean percibidos como diferentes por un mínimo de 41 y un máximo de 59 sujetos.

7.10. Intervalo de confianza para el parámetro *proporción* con $n = 15$ y $P_A = 0{,}50$.

- $\alpha = 0{,}05$.
- $|Z_{\alpha/2}| = |Z_{0{,}025}| = 1{,}96$.
- $\sigma_{P_A} = \sqrt{P_A(1-P_A)/n} = \sqrt{0{,}50(1-0{,}50)/15} = 0{,}129$.
- $E_{\text{máx}} = 1{,}96(0{,}129) = 0{,}25$.
- $L_i = 0{,}50 - 0{,}25 = 0{,}25$.

  $L_s = 0{,}50 + 0{,}25 = 0{,}75$.

Si un sujeto responde al azar, cabe esperar que no acierte menos del 25 % ni más del 75 % de las preguntas. Es decir, no menos de 4 preguntas ni más de 11 (se obtiene el mismo resultado utilizando las probabilidades exactas de la distribución binomial).

7.11. Intervalo de confianza para el parámetro *proporción* con $n = 300$ y $P_D = 0{,}63/300 = 0{,}21$.

- $\alpha = 0{,}05$.
- $|Z_{\alpha/2}| = |Z_{0{,}025}| = 1{,}96$.
- $\sigma_{P_D} = \sqrt{P_D(1-P_D)/n} = \sqrt{0{,}21(1-0{,}21)/300} = 0{,}0235$.
- $E_{\text{máx}} = 1{,}96(0{,}0235) = 0{,}046$.
- $L_i = 0{,}21 - 0{,}046 = 0{,}164$.

  $L_s = 0{,}50 + 0{,}046 = 0{,}546$.

El psicólogo tiene razón: el porcentaje inicialmente informado parece demasiado alto. El porcentaje actual estimamos que se encuentra entre el 16,4% y el 54,6%.

7.12. Sólo es verdadera la alternativa *c* (que describe correctamente el significado de un intervalo de confianza). Un intervalo de confianza sobre la media se refiere siempre al parámetro μ; esto descarta la alternativa *a*. Cuando se construye un intervalo ya no existen variables (todo en él son constantes); por tanto, no tiene sentido hablar de probabilidad; esto descarta la alternativa *b*. En la alternativa *d* se confunde un intervalo con el 5% de los intervalos. El intervalo de confianza no se refiere a puntuaciones individuales, sino a promedios; esto descarta la alternativa *e*.

7.13. La anchura o amplitud de un intervalo de confianza depende del tamaño del error máximo ($E_{máx}$), y éste depende de dos cosas: el nivel de confianza (cuanto mayor es el nivel de confianza, más ancho es el intervalo; esto descarta la alternativa *e*) y del error típico del estadístico utilizado como estimador (cuanto mayor es el error típico, más ancho es el intervalo; esto descarta la alternativa *f*). Ahora bien, el error típico depende a su vez de la dispersión de la población (cuanto mayor es la dispersión poblacional, más ancho es el intervalo; esto descarta la alternativa *c*) y del tamaño muestral (cuanto mayor es el tamaño muestral, menor es el error típico; esto descarta las alternativas *a* y *b*). Sólo es correcta, por tanto, la alternativa *d*.

7.14. La interpretación es incorrecta. Un intervalo de confianza nunca se refiere a las puntuaciones individuales de los sujetos, sino a algún parámetro (en este caso, al cociente intelectual medio de los universitarios españoles). La interpretación correcta de este intervalo es que tenemos una confianza del 95% en que el verdadero cociente intelectual medio de los estudiantes universitarios se encuentre dentro del intervalo construido (98, 106).

8

Introducción a la inferencia estadística II. El contraste de hipótesis

La *estimación de parámetros* estudiada en el capítulo anterior es una de las dos estrategias propuestas para realizar inferencias. Ya hemos señalado que la otra estrategia es el *contraste de hipótesis*. Un **contraste de hipótesis** (*hypothesis test*), también llamado **contraste** o **prueba de significación** (*significance test*), es una estrategia diseñada para *tomar decisiones*: un procedimiento que permite decidir si una proposición acerca de una población puede ser mantenida o debe ser rechazada.

En la investigación empírica es frecuente encontrarse con *problemas de conocimiento* surgidos a partir de conocimientos ya existentes o a partir de la observación de nuevos datos: ¿es el tratamiento *A* más apropiado que el *B* para aliviar los síntomas de los pacientes con trastorno depresivo?, ¿son los sujetos que se sienten inseguros más agresivos que los que se sienten seguros?, ¿difieren los hombres y las mujeres en intención de voto?, etc. Estos interrogantes son sólo un pequeño ejemplo de la multitud de *problemas* que se generan en la investigación empírica en las ciencias sociales y de la salud. Tales interrogantes surgen, en general, en el seno de una teoría que intenta dar cuenta de alguna parcela de la realidad y se plantean con la intención de cubrir alguna laguna concreta de conocimiento que esa teoría no cubre o para corroborar una parte o el total de esa teoría.

Surgido el problema, el paso siguiente consiste en aventurar algún tipo de solución al mismo. Esta solución tentativa o provisional suele tomar forma de afirmación directamente verificable (es decir, empíricamente contrastable; de no ser así, nos moveríamos en el terreno de la especulación y no en el de la ciencia) en la que se establece de forma operativa el comportamiento de la variable o variables involucradas en el problema. Esa afirmación verificable recibe el nombre de **hipótesis científica** o *de investigación* (ver Pereda, 1987, Capítulo 5). Así, ante la pregunta (problema de conocimiento) *¿cuál de dos tratamientos, A y B, es más eficaz para aliviar los síntomas de pacientes con depresión?*, podría aventurarse la hipótesis de que *los tratamientos son igualmente eficaces*. Por supuesto, habría que definir con precisión (operativamente) en qué consiste cada tratamiento, y qué se entiende por depresión y cómo medirla. Sólo entonces la afirmación podría considerarse una hipótesis científica.

Hecho esto, ya se estaría en condiciones de iniciar el proceso de verificación de esa hipótesis. Y el proceso de verificación habitualmente utilizado en las ciencias empíricas sigue los pasos que en este apartado se describen bajo la denominación de *contraste de hipótesis*.

El contraste de hipótesis

El primer paso del proceso de verificación de una hipótesis[1] consiste en *formular estadísticamente la hipótesis científica que se desea contrastar*; es decir, en transformar la hipótesis científica en **hipótesis estadística**. Esto supone que de una hipótesis científica se derivan una serie de consecuencias referidas a la forma de una o varias distribuciones poblacionales, o al valor de uno o más parámetros de esas distribuciones. Así, por ejemplo, la hipótesis científica *los tratamientos antidepresivos A y B son igualmente eficaces* implica, en términos estadísticos, $\mu_A = \mu_B$; es decir, que el promedio μ_A de la distribución de la variable *depresión* en la población de pacientes que reciben el tratamiento *A* es igual al promedio μ_B de esa misma distribución en la población de pacientes que reciben el tratamiento *B*.

Formulada la hipótesis estadística, el segundo paso de un contraste de hipótesis consiste en buscar **evidencia empírica** *relevante capaz de informar sobre si la hipótesis formulada es o no sostenible*. Esto, en general, no resulta demasiado complicado de conseguir: parece razonable pensar que, si una hipótesis concreta referida a una distribución poblacional es *cierta*, al extraer una muestra de esa población debe encontrarse un resultado muestral similar al que esa hipótesis propone para la distribución poblacional. Si una hipótesis afirma que *dos tratamientos antidepresivos son igualmente eficaces* ($\mu_A = \mu_B$) y se asume que esa hipótesis es cierta, debe esperarse que, al extraer una muestra aleatoria de la población de pacientes tratados con *A* y otra de la población de pacientes tratados con *B*, el nivel de depresión observado en ambas muestras, $\bar{Y}_A$ y $\bar{Y}_B$, sea similar.

Ahora bien, si el resultado muestral encontrado no coincide con la afirmación establecida en la hipótesis, pueden estar ocurriendo dos cosas diferentes: bien la hipótesis planteada no es cierta y, por tanto, es incapaz de ofrecer predicciones correctas; bien la hipótesis es *cierta* y la discrepancia observada entre la hipótesis y los datos es sólo producto de las fluctuaciones propias del azar muestral. La clave está precisamente en poder discernir cuándo una discrepancia entre el comportamiento de los datos y la afirmación establecida en la hipótesis es lo bastante grande como para poder considerar que el resultado muestral observado es incompatible con la hipótesis formulada, es decir, lo bastante grande como para considerar que la discrepancia hipótesis-datos no es explicable por las fluctuaciones del azar muestral sino por el hecho de que la hipótesis planteada es falsa.

Hace falta, y éste es el tercer paso del proceso, una **regla de decisión**. Y esa regla se establece en términos de *probabilidad*. Si en el ejemplo propuesto sobre la eficacia de dos tratamientos se pudiera trabajar con las poblaciones completas de pacientes tratados con *A* y con *B* (es decir, si se pudiera medir el nivel de depresión en *todos* los pacientes tratados con *A* y con *B*), no habría que recurrir a la teoría de la probabilidad porque tampoco sería necesario efectuar ningún tipo de contraste de hipótesis: se conocerían los valores poblacionales μ_A y μ_B, y se sabría si son iguales o no. Pero el hecho de tener que trabajar con *muestras* en lugar

[1] Por supuesto, no todas las hipótesis científicas requieren de la utilización del contraste de hipótesis para ser verificadas. Recordemos a este respecto lo dicho en el primer capítulo sobre los fenómenos deterministas y aleatorios. Una afirmación del tipo “tal persona posee una inteligencia superior a la media” puede verificarse simplemente observando a ese sujeto. Sin embargo, una afirmación del tipo “las personas autoritarias tienen una inteligencia inferior a la media” no puede ser verificada recurriendo sólo a la observación: difícilmente alguien podría observar a *todas* las personas autoritarias. Es justamente en las situaciones en las que no se tiene acceso a todos los elementos de la población donde las estrategias basadas en la inferencia estadística (la estimación de parámetros y el contraste de hipótesis) aportan todo su potencial para la verificación de hipótesis científicas.

de *poblaciones* obliga a recurrir a la inferencia y a tener que establecer una regla de decisión en términos de probabilidad.

Ahora bien, el número de reglas de decisión que se pueden establecer en una situación particular es casi ilimitado. Por supuesto, unas reglas serán mejores o más útiles que otras y, probablemente, ninguna de ellas será lo bastante buena como para resultar útil en todo tipo de situaciones. Afortunadamente, la *teoría de la decisión* se ha encargado de elaborar algunos argumentos que pueden trasladarse al contexto del contraste de hipótesis. En general, la regla de decisión que se utiliza en los contrastes de hipótesis se basa en el siguiente razonamiento: si, suponiendo cierta la hipótesis, el resultado muestral observado es *improbable*, se considerará que la hipótesis es incompatible con los datos; por el contrario, si, suponiendo cierta la hipótesis, el resultado muestral observado es probable, se considerará que la hipótesis es compatible con los datos[2]. Por tanto, se trata de una regla de decisión que se basa en el grado de compatibilidad (expresada ésta en términos de probabilidad) existente entre la *hipótesis* y los *datos*.

Imaginemos que estamos interesados en averiguar si un grafólogo posee o no la capacidad de detectar la presencia de trastornos depresivos a partir de la escritura. Para ello, podríamos comenzar formulando la hipótesis de que *el grafólogo no posee tal capacidad*. Si esta hipótesis es cierta, al presentar al grafólogo un par de muestras de escritura (una perteneciente a un paciente con trastorno y otra a un paciente sin trastorno) para que elija la que pertenece al paciente con trastorno, cabe esperar que responda al azar (se está asumiendo que la hipótesis es cierta), por lo que la probabilidad de acierto será de 0,5. Por el contrario, si la hipótesis es falsa (y, por tanto, el grafólogo sí posee la mencionada capacidad), al presentarle el mismo par de muestras de escritura, la probabilidad de acierto será mayor que 0,5, es decir, mayor que la probabilidad de acertar por azar. En este escenario, la hipótesis de que *el grafólogo no posee la capacidad de diagnosticar trastornos depresivos a través de la escritura* implica una afirmación estadística del tipo

$$\pi_{\text{acierto}} \leq 0{,}5$$

Para contrastar esta hipótesis se pueden presentar, en lugar de un par de muestras de escritura, 10 pares. Si la hipótesis es verdadera, cabe esperar encontrar no más de 5 aciertos (es decir, no más del número de aciertos esperable por azar). Por el contario, si la hipótesis es falsa, cabe esperar encontrar un número de aciertos superior a 5 (es decir, más aciertos de los esperables por azar).

Ahora bien, si el grafólogo obtiene 6 aciertos, ¿podrá decirse que ese resultado es mayor que el esperable por azar? ¿Y si obtiene 7? ¿Con cuántos aciertos podremos decir que el grafólogo ha superado el resultado más alto esperable sólo por azar? Para responder a esta pregunta, en lugar de basarnos en nuestras apreciaciones subjetivas, recurrimos a la teoría de la probabilidad intentando establecer una regla que nos permita decidir cuándo los datos son compatibles con esa hipótesis y cuándo no lo son. Aplicando esta regla, un número de aciertos esperable por azar (es decir, un resultado *probable* cuando se diagnostica al azar) llevará a decidir que la hipótesis planteada es compatible con los datos y a sospechar que el grafólogo no posee la capacidad de diagnosticar a partir de la escritura; por el contrario, un número de aciertos mayor que el esperable por azar (es decir, un resultado *improbable* cuando se diagnostica al azar) llevará a decidir que la hipótesis planteada es incompatible con los datos y a

[2] Sobre el significado de los términos "probable" e "improbable" volveremos más adelante.

concluir que el grafólogo sí posee esa capacidad (pues si "$\pi_{\text{acierto}} \leq 0{,}5$" es una afirmación incorrecta, entonces la afirmación correcta debe ser $\pi_{\text{acierto}} > 0{,}5$). Así pues, resumiendo:

> Un **contraste de hipótesis** es un *proceso de decisión en el que una hipótesis formulada en términos estadísticos es puesta en relación con los datos empíricos para determinar si es o no compatible con ellos*.

Y, en cuanto proceso que es, se desarrolla en una serie de pasos que se ajustan a la lógica que acabamos de resumir y que se describe con más detalle a continuación[3].

Las hipótesis estadísticas

Una hipótesis estadística es una afirmación sobre una o más distribuciones de probabilidad; más concretamente, sobre la *forma* de una o más distribuciones de probabilidad, o sobre el valor de uno o más *parámetros* de esas distribuciones. Las hipótesis estadísticas se suelen representar por la letra *H* seguida de una afirmación que da contenido a la hipótesis:

H: la variable *Y* se distribuye normalmente con $\mu_Y = 100$ y $\sigma_Y = 15$

H: $\pi_{\text{acierto}} = 0{,}5$

H: $\mu_Y \leq 30$

H: $Mdn_1 \neq Mdn_2$

H: $\mu_1 = \mu_2 = \mu_3 = \mu_4$

En general, una hipótesis estadística surge a partir de una hipótesis científica. Pero entre una hipótesis científica y una hipótesis estadística no existe una correspondencia exacta. La primera actúa como referente para la formulación de la segunda, pero no son la misma cosa. Una hipótesis científica se refiere a algún aspecto de la realidad; una hipótesis estadística se refiere a algún aspecto de una distribución de probabilidad. Esto significa, por ejemplo, que la expresión $\mu_A = \mu_B$ propuesta anteriormente no es la única formulación estadística que puede derivarse de la hipótesis científica *ambos tratamientos son igualmente eficaces*. En lugar del promedio μ podría utilizarse el promedio *Mdn* y establecer esta otra formulación estadística: $Mdn_A = Mdn_B$. Y todavía podría transformarse esa hipótesis científica en estadística utilizando otras estrategias muy distintas; por ejemplo: $F(Y_A) = F(Y_B)$, es decir, la función de distribución de la variable *Y* = «nivel de depresión» es la misma en la población de pacientes tratados con *A* y en la de pacientes tratados con *B*.

Existen, por tanto, varias formas distintas de expresar estadísticamente una hipótesis científica concreta. A lo largo de este capítulo y de los que siguen iremos viendo qué hipótesis estadísticas es posible plantear y cómo deben plantearse. De momento, basta con saber que el primer paso de todo contraste de hipótesis consiste en formular en términos estadísticos la afirmación contenida en la hipótesis científica que se desea contrastar.

---

[3] En la inferencia estadística no existe un único punto de vista. Es frecuente encontrarse con la distinción entre el *enfoque clásico*, en el que se considera que la única información disponible sobre la población es la que contienen las muestras, y el *enfoque bayesiano*, en el que, además de la información muestral, se hace uso de conocimientos previos. Las ideas sobre el contraste de hipótesis tal como se expone aquí fueron introducidas inicialmente por Ronald A. Fisher (1925, 1955) y completadas por Neyman y Pearson (1928, 1932, 1933), aunque las principales aportaciones de Neyman y Pearson las dejaremos para el segundo volumen. Esta versión del contrate de hipótesis debe ser enmarcada dentro del enfoque clásico.

Dicho esto, es necesario señalar que, aunque hasta ahora se han venido proponiendo ejemplos en los que se ha formulado una sola hipótesis, lo cierto es que los contrastes de hipótesis se basan en la formulación de dos hipótesis:

1. La *hipótesis nula*, representada por H_0.
2. La *hipótesis alternativa*, representada por H_1.

La **hipótesis nula** H_0 es la hipótesis que se pone a prueba. Es la clave de todo el proceso: no sólo es el punto de partida sino que, según veremos, guía todo el proceso y es sobre quien finalmente recae la decisión (de hecho, a los contrastes de hipótesis también se les suele llamar *pruebas de significación de la hipótesis nula*). Esta hipótesis adopta la forma de afirmación concreta sobre la forma de una distribución de probabilidad o sobre el valor de alguno de los parámetros de esa distribución:

H_0: La variable Y se distribuye normalmente con $\mu_Y = 100$ y $\sigma_Y = 15$.
H_0: $\pi_1 = \pi_2$.
H_0: $\mu_1 = \mu_2$.
H_0: $\rho_{XY} = 0$.
H_0: $\pi_{\text{acierto}} = 0{,}5$.

La **hipótesis alternativa** H_1 es la negación de la hipótesis nula. H_1 incluye todo lo que H_0 excluye. Mientras que H_0 es una hipótesis *exacta* (tal cosa es *igual* a tal otra), H_1 es *inexacta* (tal cosa es *distinta*, *mayor* o *menor* que tal otra):

H_1: La variable Y no se distribuye normalmente con $\mu_Y = 100$ y $\sigma_Y = 15$.
H_1: $\pi_1 > \pi_2$.
H_1: $\mu_1 < \mu_2$.
H_1: $\rho_{XY} \neq 0$.
H_1: $\pi_{\text{acierto}} < 0{,}5$.

Cuando en H_1 aparece el signo *distinto* ($\neq$), se dice que el contraste es *bilateral* o bidireccional. Cuando en H_1 aparece el signo *menor que* ($<$) o *mayor que* ($>$) se dice que el contraste es *unilateral* o unidireccional. Enseguida volveremos sobre esta distinción.

La hipótesis nula y la hipótesis alternativa suelen plantearse como hipótesis rivales. Son hipótesis exhaustivas (agotan todas las posibilidades) y mutuamente exclusivas (no se solapan), lo cual implica que si una es verdadera, la otra es necesariamente falsa. Según esto, en los ejemplos propuestos anteriormente pueden plantearse las siguientes hipótesis:

a) H_0: $\mu_A = \mu_B$.
H_1: $\mu_A \neq \mu_B$.

b) H_0: $\pi_{\text{acierto}} \leq 0{,}5$.
H_1: $\pi_{\text{acierto}} > 0{,}5$.

Las hipótesis del párrafo *a* se refieren al ejemplo sobre la eficacia de dos tratamientos antidepresivos: la hipótesis nula afirma que el nivel medio de depresión de los pacientes tratados con *A* es el mismo que el de los tratados con *B*; la hipótesis alternativa afirma que no es el

mismo. Las hipótesis del párrafo *b* se refieren al ejemplo del grafólogo capaz de diagnosticar trastornos depresivos a través de la escritura: la hipótesis nula afirma que el grafólogo no posee tal capacidad; la hipótesis alternativa afirma que sí la posee. En ambos casos, H_0 y H_1 se plantean como hipótesis exhaustivas y mutuamente exclusivas.

Conviene no pasar por alto un detalle de especial importancia: el signo *igual* (=), tanto si va solo ($\mu_A = \mu_B$) como si va acompañado ($\pi_{\text{acierto}} \leq 0{,}5$), *siempre* va en la hipótesis nula. Ya hemos señalado que H_0 es la hipótesis que se somete a contraste. Esto significa que la afirmación *concreta* establecida en H_0 (y la única afirmación concreta que se está haciendo es la que corresponde al signo "=") es el punto de partida de todo el proceso. Es decir, tanto si H_0 es *exacta* ($\mu_A = \mu_B$) como si es *inexacta* ($\pi_{\text{acierto}} \leq 0{,}5$), todo el proceso de decisión va a estar basado en un modelo probabilístico construido a partir de la afirmación concreta correspondiente al signo "=" presente en H_0. Enseguida veremos que es justamente ese modelo probabilístico el que ofrece la información necesaria para tomar una decisión sobre H_0.

Los supuestos del contraste

Para que una hipótesis estadística pueda predecir un resultado muestral con cierta exactitud es necesario que *la distribución muestral con la que se va a trabajar esté completamente especificada*.

Por ejemplo, una afirmación del tipo "la variable *Y* se distribuye normalmente con parámetros $\mu_Y = 100$ y $\sigma_Y = 15$" es una afirmación que identifica por completo las características de la población de referencia: define una población normal con parámetros conocidos. Sabemos qué se puede esperar de una muestra aleatoriamente seleccionada de esa población. Pero una afirmación del tipo "$\mu_Y = 30$" hace referencia a una población de la que únicamente se sabe que la media de la variable *Y* vale 30. No sabemos qué se puede esperar en una muestra seleccionada de esa población a no ser que impongamos alguna condición extra como la forma de la población y el valor de la desviación típica.

Con las hipótesis de un contraste pasa algo parecido. Enseguida veremos que la distribución que es necesario conocer en un contrate de hipótesis es la distribución muestral del estadístico del contraste. Y, como las hipótesis nulas no siempre permiten especificar por completo esa distribución muestral, es necesario establecer condiciones adicionales a las que impone la hipótesis nula. Estas condiciones adicionales son las que llamamos *supuestos* del contraste.

En el ejemplo del grafólogo supuestamente capaz de detectar trastornos depresivos a través de la escritura, para verificar si el grafólogo posee o no esa capacidad, se han planteado las hipótesis estadísticas H_0: $\pi_{\text{acierto}} \leq 0{,}5$ y H_1: $\pi_{\text{acierto}} > 0{,}5$. Y para contrastar esas hipótesis se han presentado al grafólogo 10 pares de muestras de escritura. Pues bien, si los 10 pares de muestras de escritura se presentan de forma independiente (muestra aleatoria) y en cada presentación sólo hay dos resultados posibles (acierto-error) con $\pi_{\text{acierto}} = 0{,}5$ en cada presentación, la variable *número de aciertos* tendrá una distribución de probabilidad completamente especificada (la binomial, con parámetros $n = 10$ y $\pi_{\text{acierto}} = 0{,}5$; ver, en el Capítulo 6, el apartado *Distribución muestral del estadístico proporción*) y eso permitirá poder tomar una decisión respecto a H_0 en términos de probabilidad.

Normalmente será necesario establecer supuestos sobre las características de las poblaciones muestreadas. Pero también será necesario establecer otro tipo de supuestos relacionados

con la forma de llevar a cabo el estudio: *si la muestra es aleatoria..., si las presentaciones son independientes...*

Resumiendo: los supuestos de un contraste de hipótesis son el conjunto de condiciones que deben darse (forma de la población de partida, características de la muestra elegida, etc.) para que la distribución de probabilidad en la que se basará la decisión sobre H_0 quede completamente especificada.

El estadístico del contraste y su distribución muestral

El estadístico del contraste es un valor muestral que cumple una doble condición: (1) ofrece información relevante sobre la afirmación establecida en la hipótesis nula y (2) tiene distribución muestral conocida.

Si la hipótesis que se desea contrastar es H_0: $\mu_Y = 30$, debe recurrirse a un estadístico capaz de detectar cualquier desviación empírica de la afirmación establecida en H_0. Obviamente, ni S_Y, ni $P_{acierto}$, por citar algunos estadísticos conocidos, ofrecerán información relevante sobre el parámetro μ_Y. Para contrastar H_0: $\mu_Y = 30$, lo razonable será utilizar la información muestral basada en el estadístico $\overline{Y}$; del mismo modo, si la hipótesis que se desea contrastar es del tipo H_0: $\pi_{acierto} = 0{,}70$, lo razonable será recurrir a un estadístico que pueda ofrecer información relevante sobre $\pi_{acierto}$, por ejemplo, $P_{acierto}$; etc.

La segunda condición que debe cumplir un estadístico para que pueda utilizarse como estadístico del contraste es la de tener *distribución muestral conocida*. Precisamente la distribución muestral del estadístico del contraste es la que contiene las probabilidades en las que más tarde se basará la decisión sobre H_0.

Por tanto, una vez formuladas las hipótesis, el siguiente paso consiste en seleccionar un estadístico capaz de informar sobre ellas y en fijar las condiciones (supuestos) necesarias para conseguir determinar su distribución muestral. En el ejemplo sobre el grafólogo supuestamente capaz de diagnosticar trastornos depresivos a partir de la escritura se habían planteado las hipótesis H_0: $\pi_{acierto} \leq 0{,}5$ y H_1: $\pi_{acierto} > 0{,}5$. Existen dos estadísticos (en realidad los dos son el mismo, pues uno es transformación lineal del otro) que aportan información relevante sobre esa hipótesis:

$n_{acierto}$ = «número de aciertos o de diagnósticos correctos».
$P_{acierto}$ = «proporción de aciertos o de diagnósticos correctos».

Asumiendo, según se ha señalado antes, que las presentaciones de los 10 pares de muestras de escritura son independientes entre sí y que la probabilidad de cada uno de los dos resultados posibles (acierto-error) es la misma en cada presentación, la distribución muestral de las variables (estadísticos del contraste) $n_{acierto}$ y $P_{acierto}$ es la binomial con parámetros $n = 10$ y $\pi_{acierto} = 0{,}5$. Por tanto, la probabilidad asociada a cada uno de los valores de $n_{acierto}$ y $P_{acierto}$ puede obtenerse aplicando la ecuación de la función binomial (ver ecuación [3.4]) o, más rápido y sencillo, recurriendo a la Tabla B del Apéndice final (sin olvidar que la tabla ofrece probabilidades acumuladas).

La Tabla 8.1 recoge estas probabilidades, es decir, las distribuciones muestrales de $n_{acierto}$ y $P_{acierto}$ con $n = 10$ y $\pi_{acierto} = 0{,}5$. En la tabla puede comprobarse, por ejemplo, que la probabilidad de encontrar 10 aciertos (es decir, la probabilidad de $n_{acierto} = 10$ o $P_{acierto} = 1$) vale 0,001. Y también puede comprobarse, por ejemplo, que la probabilidad de encontrar 9 aciertos o más

(es decir, la probabilidad de $n_{acierto} \geq 9$ o $P_{acierto} \geq 0{,}9$), siempre asumiendo $\pi_{acierto} = 0{,}5$, vale 0,010 + 0,001 = 0,011. Según veremos enseguida, es justamente en estas probabilidades en las que se basará la decisión sobre H_0.

Por tanto, los estadísticos $n_{acierto}$ y $P_{acierto}$ permiten contrastar la hipótesis H_0: $\pi_{acierto} \leq 0{,}5$ porque ambos cumplen las dos condiciones exigidas a los estadísticos utilizados en los contrastes de hipótesis: (1) contienen información relevante sobre H_0 y (2) tienen distribución muestra conocida. Por tanto, al contrastar H_0: $\pi_{acierto} \leq 0{,}5$, tanto $n_{acierto}$ como $P_{acierto}$ sirven como estadísticos del contraste.

Tabla 8.1. Distribución muestral de $n_{acierto}$ y $P_{acierto}$ (con $n = 10$ y $\pi_{acierto} = 0{,}5$)

$n_{acierto}$	$P_{acierto}$	$f(n_{acierto}) = f(P_{acierto})$
0	0,0	0,001
1	0,1	0,010
2	0,2	0,044
3	0,3	0,117
4	0,4	0,205
5	0,5	0,246
6	0,6	0,205
7	0,7	0,117
8	0,8	0,044
9	0,9	0,010
10	1,0	0,001

La regla de decisión

La regla de decisión es el criterio que se utiliza para decidir si la hipótesis nula puede mantenerse o debe rechazarse. La lógica en la que se basa esta regla es bastante simple: la hipótesis nula se mantiene o rechaza dependiendo de su *grado de compatibilidad con los datos*; y los datos están resumidos en el estadístico del contraste. Para determinar el grado de compatibilidad entre la hipótesis nula y los datos, la distribución muestral del estadístico del contraste se divide en dos zonas exclusivas y exhaustivas: la *zona de rechazo* y la *zona de aceptación*.

La **zona de rechazo**, también llamada **zona crítica**, es la zona de la distribución muestral correspondiente a los valores del estadístico del contraste que se encuentran tan alejados de la afirmación establecida en H_0 que es muy poco probable que ocurran si H_0, como se supone, es verdadera; es decir, la zona en la que se encuentran los datos *poco compatibles* con H_0. La probabilidad asociada a esta zona de rechazo o crítica se denomina **nivel de significación** o **nivel de riesgo** y se representa con la letra griega α.

La **zona de aceptación** es la zona de la distribución muestral correspondiente a los valores del estadístico del contraste próximos a la afirmación establecida en H_0. Es, por tanto, la zona en la que se encuentran los valores del estadístico que es probable que ocurran si H_0, como se supone, es verdadera; es decir, la zona en la que se encuentran los datos *compatibles* con H_0. La probabilidad asociada a esta zona de aceptación se denomina **nivel de confianza** y se representa mediante $1 - \alpha$.

Una vez definidas las zonas de rechazo y de aceptación, se aplica la siguiente **regla de decisión**:

> *Rechazar H_0 cuando el estadístico del contraste toma un valor perteneciente a la zona de rechazo o crítica; mantener H_0 cuando el estadístico del contraste toma un valor perteneciente a la zona de aceptación.*

Por tanto, se rechaza una H_0 particular *porque* eso significa que el valor del estadístico del contraste se aleja demasiado de la predicción establecida en esa hipótesis, es decir, *porque*, si esa H_0 fuera verdadera, el estadístico del contraste no debería tomar ese valor (sería muy poco probable que lo tomara); si de hecho lo toma, la conclusión razonable es que esa H_0 no puede ser verdadera (es importante advertir que la decisión siempre se toma sobre H_0).

Con esta regla de decisión se está asumiendo que la probabilidad asociada al estadístico del contraste indica el grado de compatibilidad existente entre la hipótesis nula y los datos. Esta probabilidad (grado de compatibilidad) recibe el nombre de **nivel crítico** (también se le llama *nivel de significación observado*) y se representa mediante p:

$$p = \textit{nivel crítico} = P(D \mid H_0) \qquad (D = \text{Datos}) \qquad [8.1]$$

Es decir, p representa la probabilidad de encontrar, en la distribución muestral definida por H_0, los datos de hecho encontrados (más concretamente, la probabilidad de encontrar datos tan alejados, o más alejados, de la afirmación establecida en H_0 como los de hecho encontrados). Aplicando este criterio de compatibilidad entre la hipótesis nula y los datos, la regla de decisión puede formularse de esta otra manera:

> *Rechazar H_0 si $p < \alpha$; mantenerla en caso contrario*

El tamaño de las zonas de rechazo y aceptación se determina fijando el valor de α, es decir, fijando el *nivel de significación* o *nivel de riesgo* con el que se desea trabajar. Por supuesto, si se tiene en cuenta que α es la probabilidad que se va a considerar como lo bastante pequeña para que valores con esa probabilidad o menor no ocurran bajo H_0, se comprenderá que α será, necesariamente, un valor *pequeño*. Cómo de *pequeño* es algo que debe establecerse de forma arbitraria, si bien el valor habitualmente propuesto para α en la literatura estadística y aceptado por la comunidad científica es 0,05 (también referido como nivel de significación del 5%).

La forma de dividir la distribución muestral en zona de rechazo y zona de aceptación depende de que el contraste sea bilateral o unilateral. En un contraste **bilateral** o bidireccional no se tiene una idea previa sobre la dirección en la que pueden aparecer resultados muestrales incompatibles con H_0. Esto es lo que ocurre, por ejemplo, cuando se desea comprobar si un parámetro toma o no un determinado valor, o si dos grupos difieren en alguna variable, o si dos variables son independientes:

1. $H_0: \pi_{\text{acierto}} = 0{,}5.$
   $H_1: \pi_{\text{acierto}} \neq 0{,}5.$
2. $H_0: \mu_A = \mu_B.$
   $H_1: \mu_A \neq \mu_B.$
3. $H_0: \rho_{XY} = 0.$
   $H_1: \rho_{XY} \neq 0.$

En el caso 1, H_0 será rechazada tanto si π_{acierto} es mayor que 0,5 como si es menor; en el caso 2, H_0 será rechazada tanto si μ_A es mayor que μ_B como si μ_A es menor que μ_B; en el caso 3, H_0 será rechazada tanto si la relación es positiva como si es negativa. Todos estos contrastes son bilaterales: las hipótesis alternativas no indican la dirección en la que se encuentran los resultados muestrales incompatibles con H_0 (lo cual se expresa con el signo "≠").

En un contraste **unilateral** o unidireccional se tiene una idea previa sobre la dirección en la que se encuentran los resultados muestrales incompatibles con H_0. Esto es lo que ocurre, por ejemplo, cuando se desea comprobar si el valor de un parámetro ha aumentado, o si un grupo supera a otro en alguna variable, o si dos variables se encuentran positivamente relacionadas:

1. $H_0: \pi_{\text{acierto}} \leq 0{,}50$.
   $H_1: \pi_{\text{acierto}} > 0{,}50$.
2. $H_0: \mu_A \leq \mu_B$.
   $H_1: \mu_A > \mu_B$.
3. $H_0: \rho_{XY} \leq 0$.
   $H_1: \rho_{XY} > 0$.

En el caso 1, H_0 será rechazada si π_{acierto} es mayor que 0,50, pero no si es menor; en el caso 2, H_0 será rechazada si μ_A es mayor que μ_B, pero no si es menor; en el caso 3, H_0 será rechazada si la relación es positiva, pero no si es negativa. Todos estos contrastes son unilaterales: las hipótesis alternativas contienen una predicción concreta (expresada con los signos "<" y ">") sobre la dirección en la que se encuentran los resultados muestrales incompatibles con la afirmación establecida en H_0.

La zona de rechazo o crítica, por tanto, debe situarse allí donde pueden aparecer los valores muestrales incompatibles con H_0, es decir, allí donde indica H_1. Y esto es algo que depende únicamente de lo que interese estudiar en cada caso concreto. Por ejemplo, para comparar la eficacia de dos tratamientos sin una expectativa justificada (estudios previos, intereses concretos, etc.) sobre cuál de los dos es más eficaz, lo razonable es plantear un contraste bilateral (H_1: $\mu_A \neq \mu_B$). Lo cual significa que la zona crítica debe recoger los valores muestrales que vayan tanto en la dirección $\mu_A - \mu_B > 0$ como en la dirección $\mu_A - \mu_B < 0$. Dicho de otro modo, si H_0: $\mu_A = \mu_B$ es falsa, lo será tanto si μ_A es mayor que μ_B como si μ_A es menor que μ_B; y la zona crítica deberá recoger ambas posibilidades[4]. Por esta razón, *en los contrastes bilaterales, la zona crítica se encuentra repartida*[5]*, generalmente a partes iguales, entre las dos colas de la distribución muestral* (Figura 8.1, izquierda).

Sin embargo, para comprobar si un grafólogo posee o no la capacidad de diagnosticar trastornos depresivos a partir de la escritura, lo razonable es plantear un contraste unilateral (H_1: $\pi_{\text{acierto}} > 0{,}5$), pues sólo tiene sentido considerar que el grafólogo posee tal capacidad si

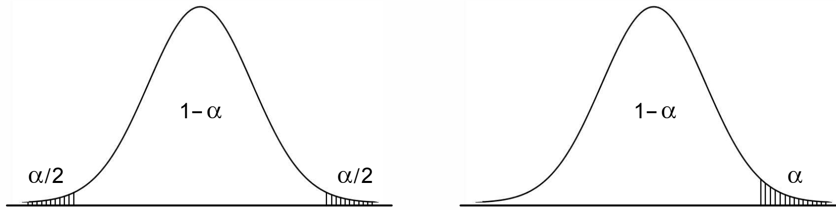
---

[4] Por supuesto, si se desea contrastar, no si dos tratamientos difieren, sino si uno es mejor que el otro, habrá que plantear un contraste unilateral.

[5] Existen excepciones a esta regla. Dependiendo del estadístico utilizado y de su distribución muestral, puede ocurrir que la zona crítica de un contraste bilateral esté, toda ella, ubicada en la cola derecha de la distribución. Cuando se utiliza la distribución normal o la distribución *t* de *Student*, la zona crítica de los contrastes bilaterales está repartida, generalmente en partes iguales, entre las dos colas de la distribución muestra. Pero la cosa cambia cuando se utilizan otras distribuciones como, por ejemplo, la distribución χ^2. Esto es algo que tendremos ocasión de estudiar más adelante.

la proporción de aciertos es *mayor* que la esperable por azar (no si esa proporción es *menor*). En este caso, los únicos valores muestrales incompatibles con H_0 son los que van en la dirección $\pi_{acierto} > 0{,}5$, que es la dirección apuntada en H_1. Y la zona crítica debe reflejar esta circunstancia quedando ubicada en la cola derecha de la distribución muestral. Por tanto, *en los contrastes unilaterales, la zona crítica se encuentra en una de las dos colas de la distribución muestral* (Figura 8.1, derecha).

Figura 8.1. Ejemplo de zonas críticas en un contraste bilateral (izquierda) y unilateral derecho (derecha) en una distribución muestral de forma normal



Según esto, las reglas de decisión para los contrastes de nuestros dos ejemplos (el de las diferencias entre dos tratamientos y el del grafólogo capaz de diagnosticar trastornos depresivos a partir de la escritura) pueden concretarse de la siguiente manera:

1. Rechazar H_0: $\mu_A = \mu_B$ si el estadístico del contraste cae en la zona crítica, es decir, si toma un valor mayor que el cuantil $100(1 - \alpha/2)$ o menor que el cuantil $100(\alpha/2)$ de su distribución muestral.

   O bien: rechazar H_0: $\mu_A = \mu_B$ si el estadístico del contraste toma un valor tan grande o tan pequeño que la probabilidad de obtener un valor tan extremo o más que el obtenido es menor que $\alpha/2$. Es decir, rechazar H_0 si $p/2 < \alpha/2$; o, lo que es lo mismo, si $p < \alpha$.

2. Rechazar H_0: $\pi_{acierto} \leq 0{,}5$ si el estadístico del contraste cae en la zona crítica, es decir, si toma un valor mayor que el percentil $100(1 - \alpha)$ de su distribución muestral.

   O bien: rechazar H_0: $\pi_{acierto} \leq 0{,}5$ si el estadístico del contraste toma un valor tan grande que la probabilidad de obtener un valor como ése o mayor es menor que α. Es decir, rechazar H_0 si $p < \alpha$.

La regla de decisión encierra un argumento claro acerca del rol que juega el azar muestral en la variabilidad observada en los datos. Cuando se decide *no rechazar* una H_0 se está asumiendo que el efecto (diferencia, relación) observado puede explicarse sin recurrir a otra cosa que a la variabilidad propia del azar muestral; cuando se decide *rechazar* una H_0 se está descartando el azar muestral como única explicación del efecto observado. Así, cuando un efecto no es lo bastante grande como para decidir rechazar H_0, lo que se está queriendo decir es que ese efecto se encuentra dentro del rango de valores esperables por azar si H_0 se asume verdadera. Por el contrario, cuando el efecto es lo bastante grande como para decidir rechazar H_0, lo que se está queriendo decir es que el efecto excede el rango de valores esperables por azar cuando H_0 es verdadera; y esto implica que el efecto observado no puede explicarse únicamente a partir de la variabilidad atribuible al azar muestral. Cuando se da esta circunstancia, se dice que el resultado es *estadísticamente significativo*.

La decisión

Planteada la hipótesis, formulados los supuestos, obtenido el estadístico del contraste y su distribución muestral, y establecida la regla de decisión, el paso siguiente de un contraste consiste en tomar una decisión. Tal decisión se toma, siempre, sobre H_0, y siempre consiste en rechazarla o mantenerla de acuerdo con las condiciones establecidas en la regla de decisión: si el estadístico del contraste cae en la zona crítica ($p < \alpha$), se rechaza H_0; si el estadístico del contraste cae en la zona de aceptación ($p > \alpha$), se mantiene H_0.

La decisión, así planteada, parece no revestir problemas. Pero eso no es del todo cierto. Conviene resaltar un aspecto importante de este proceso de decisión que no siempre es tenido en cuenta. Si se *rechaza* una H_0 particular se está afirmado que ha quedado probado (con las limitaciones de un procedimiento basado en probabilidades) que esa hipótesis es falsa. Por el contrario, si se mantiene, no se está afirmando que ha quedado probado que es verdadera; simplemente se está afirmando que se considera compatible con los datos, es decir, que no se dispone de evidencia empírica suficiente para rechazarla. Por tanto:

> **Mantener** una hipótesis nula significa que se considera que *esa hipótesis es compatible con los datos*. **Rechazar** una hipótesis nula significa que se considera probado (con la limitación señalada) que *esa hipótesis es falsa*.

La razón de esta asimetría en la conclusión es doble. Por un lado, dada la naturaleza inespecífica de H_1, raramente es posible afirmar que H_1 no es verdadera; las desviaciones pequeñas de H_0 forman parte de H_1, por lo que al mantener una H_0 particular, también se están manteniendo, muy probablemente, algunos valores de H_1; debe concluirse, por tanto, que se mantiene o no rechaza H_0, pero no que se acepta como verdadera. Por otro lado, en el razonamiento que lleva a tomar una decisión sobre H_0, puede reconocerse el argumento deductivo *modus tollens*, aunque de tipo probabilístico:

> Si H_0 es verdadera, entonces, muy probablemente, el estadístico del contraste T tomará un valor compatible con ella; T no toma un valor compatible con ella; luego, muy probablemente, H_0 no es verdadera.

Este argumento es impecable; nada hay en él que lo invalide desde el punto de vista lógico. Sin embargo, si una vez establecida la primera premisa se continúa de esta otra manera: "T toma un valor compatible con H_0; luego H_0, muy probablemente, es verdadera", se comete un error lógico llamado *falacia de la afirmación del consecuente*, pues T puede haber tomado un valor compatible con H_0 por razones diferentes de las contenidas en H_0.

Resumiendo

Probablemente ahora se entenderá mejor la definición propuesta para el contraste de hipótesis como *proceso de toma de decisiones en el que una afirmación sobre alguna característica poblacional* (hipótesis nula) *es puesta en relación con los datos empíricos* (resumidos en el estadístico del contraste y su distribución muestral) *para determinar si es o no compatible con ellos* (compatibilidad que se establece en términos de probabilidad: p).

Todos los contrastes de hipótesis siguen la misma lógica: hipótesis, supuestos, estadístico del contraste y distribución muestral, y decisión basada en una probabilidad que expresa el

grado de compatibilidad entre la hipótesis y los datos. Ahora bien, puesto que las situaciones concretas que interesa analizar poseen características particulares, el proceso general recién descrito necesita adaptarse a las peculiaridades de cada una de ellas. Esto es lo que hacen las técnicas de análisis que se describen en los capítulos siguientes: cada técnica (cada contraste de hipótesis o prueba de significación) es una adaptación de este proceso general a una situación concreta (situación que vendrá caracterizada por el número de variables que intervienen, la naturaleza de las variables, la forma de recoger los datos, etc.).

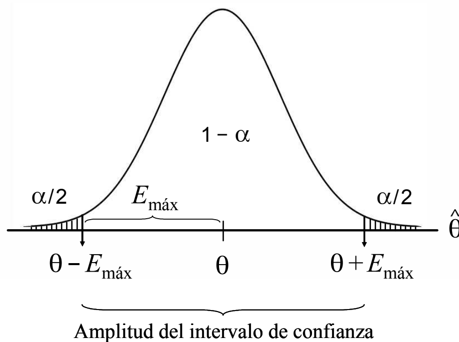
Estimación por intervalos y contraste de hipótesis

Un intervalo de confianza no sólo permite formarse una idea acerca del rango de valores entre los que cabe esperar que se encuentre el verdadero valor del parámetro estimado. También sirve para contrastar hipótesis. De hecho, la estimación por intervalos y el contraste de hipótesis se encuentran estrechamente relacionados. No en vano nos hemos referido ya a la *estimación* y al *contraste* como las dos caras de la misma moneda. Ha llegado el momento de aclarar esta relación.

Al construir un intervalo con un nivel de confianza de, por ejemplo, 0,95, se está afirmando que, de cada 100 intervalos que se construyan con muestras del mismo tamaño extraídas en las mismas condiciones, 95 de ellos incluirán el verdadero valor del parámetro estimado. Esto es equivalente a afirmar que ninguna de las hipótesis nulas referidas a los valores incluidos dentro del intervalo de confianza será rechazada en un contraste bilateral con $\alpha = 0{,}05$. En este sentido, los valores incluidos dentro del intervalo de confianza pueden interpretarse como el *conjunto de hipótesis aceptables* del correspondiente contraste bilateral; y los no incluidos, como el *conjunto de hipótesis rechazables*.

La Figura 8.2 puede ayudar a entender esta equivalencia. La curva muestra una distribución normal en la que se ha representado el tamaño del error máximo ($E_{máx}$), la amplitud del intervalo de confianza y las zonas de aceptación ($1 - \alpha$) y rechazo ($\alpha/2 + \alpha/2$) de un contraste bilateral.

Figura 8.2. Relación entre la estimación por intervalos y el contraste de hipótesis



Sabemos que cualquier intervalo construido a partir de un valor $\hat{\theta}$ de la zona rayada llevará a construir un intervalo entre cuyos límites no se encontrará el valor del parámetro θ. Tam-

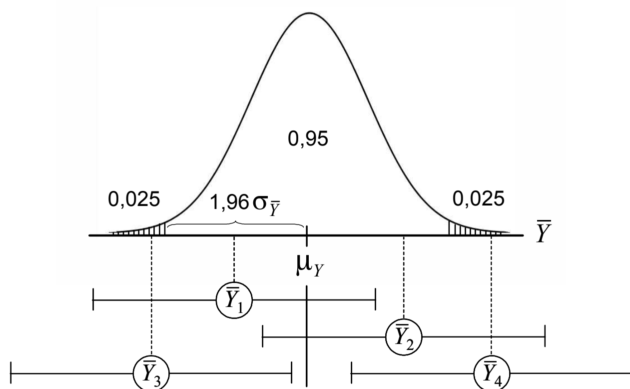
bién sabemos que cualquier valor $\hat{\theta}$ de la zona rayada que se utilice como estadístico en un contraste bilateral llevará al rechazo de H_0. Por tanto, si el valor propuesto para el parámetro θ en la H_0 de un contraste bilateral no se encuentra dentro del intervalo construido a partir de $\hat{\theta}$ con un nivel de confianza de $1 - \alpha$, entonces el contraste bilateral basado en $\hat{\theta}$ llevará al rechazo de H_0 con un nivel de significación α.

Del mismo modo, cualquier intervalo construido a partir de un valor $\hat{\theta}$ de la zona no rayada llevará no sólo a construir un intervalo entre cuyos límites se encontrará el valor del parámetro θ, sino a mantener H_0 con el correspondiente contraste bilateral. Por tanto, si el valor propuesto para el parámetro θ en H_0 en un contraste bilateral se encuentra dentro del intervalo construido a partir de $\hat{\theta}$ con un nivel de confianza de $1 - \alpha$, entonces ese contraste llevará al rechazo de H_0 con un nivel de significación α.

Supongamos que en un contraste se plantea la hipótesis nula H_0: $\mu_Y = \mu_0$ frente a la alternativa H_1: $\mu_Y \neq \mu_0$ con un nivel de confianza de 0,95. Supongamos además que, para contrastar esa hipótesis se utiliza el estadístico $\bar{Y}$, cuya distribución muestral está representada en la curva de la Figura 8.3. De acuerdo con la lógica del contraste de hipótesis recién estudiada, cualquier valor muestral $\bar{Y}$ perteneciente a la zona no rayada llevará a mantener H_0; y de acuerdo con la lógica de la estimación por intervalos estudiada en el capítulo anterior, cualquier valor muestral $\bar{Y}$ perteneciente a la zona no rayada llevará a construir un intervalo de confianza que captará el valor propuesto para μ_Y en H_0. Tal es el caso de los intervalos construidos con las medias $\bar{Y}_1$ e $\bar{Y}_2$: estas medias no sólo pertenecen a la zona de aceptación, sino que, tal como indica el gráfico, al sumarles y restarles el error máximo ($1{,}96\,\sigma_{\bar{Y}}$), se obtienen intervalos entre cuyos límites se encuentra el valor del parámetro μ_Y. Por el contrario, cualquier valor muestral $\bar{Y}$ perteneciente a la zona rayada llevará, no sólo a rechazar H_0, sino a construir un intervalo de confianza que no captará el valor propuesto para μ_Y en H_0. Tal es el caso de los intervalos construidos con las medias $\bar{Y}_3$ e $\bar{Y}_4$: estas medias no sólo pertenecen a la zona de rechazo (zona rayada), sino que al sumarles y restarles el valor del error máximo ($1{,}96\,\sigma_{\bar{Y}}$) se obtienen intervalos de confianza entre cuyos límites no se encuentra el valor del parámetro que se está intentando estimar (μ_Y).

Por tanto, al construir un intervalo de confianza para el parámetro μ_Y con un nivel de confianza de 0,95, se está asumiendo que las hipótesis nulas correspondientes a los valores de μ_Y no incluidos en ese intervalo son falsas.

Figura 8.3. Relación entre la estimación por intervalos y el contraste de hipótesis



Clasificación de los contrastes de hipótesis

Para poder aplicar un contraste de hipótesis, la primera decisión que es necesario tomar (quizá la más importante) es la de elegir correctamente el contraste concreto que permitirá poner a prueba la hipótesis que se desea contrastar. Lógicamente, si un estudio incluye varias hipótesis será necesario utilizar varios contrastes, en cuyo caso, cada uno de ellos deberá elegirse pensando en una hipótesis concreta.

Este argumento sugiere que la clasificación de los contrastes de hipótesis podría hacerse, antes que nada, tomando como referencia el tipo de hipótesis que permiten contrastar. Con este criterio, los contrastes podrían clasificarse, por ejemplo, en función de que permitan comparar medias, o comparar proporciones, o estudiar relaciones, etc. Pero lo cierto es que este criterio, por sí solo, no conduce a una clasificación del todo satisfactoria porque no resulta muy útil a quienes se inician en el análisis de datos.

Una clasificación de los contrastes de hipótesis debe servir para cubrir, al menos, estos dos objetivos: (1) ofrecer una panorámica de los contrastes disponibles y (2) ayudar al analista de datos a elegir el contraste apropiado para cada situación concreta. Creemos que ambos objetivos pueden conseguirse fácilmente si el criterio referido al tipo de hipótesis que cada contraste permite poner a prueba se complementa con otros dos: (1) el número de las variables que intervienen en el análisis y (2) la naturaleza categórica o cuantitativa de las variables analizadas[6]. El Cuadro 8.1 ofrece una clasificación de los contrastes basada en todos o parte de estos criterios. Incluye los contrastes disponibles para el análisis de una y dos variables. En este volumen estudiaremos algunos de estos contrastes (los marcados en cursiva); el resto los estudiaremos en el segundo volumen.

En la clasificación propuesta se utiliza, como primer criterio de clasificación, el *número de variables*; a continuación, la *naturaleza categórica o cuantitativa de las variables*; por último, el tipo de hipótesis que cada contraste permite poner a prueba. Creemos que, de esta manera, se facilita enormemente la elección del contraste apropiado.

Los contrastes disponibles para analizar una sola variable sirven para tomar decisiones basadas en la comparación entre un valor muestral y un valor poblacional, o entre una distribución empírica y una teórica. Para analizar **una variable categórica** se proponen dos contrastes. El primero de ellos sirve para hacer inferencias con una variable dicotómica; por ejemplo, ¿consigue un nuevo tratamiento más recuperaciones de las que se vienen obteniendo con el tratamiento convencional? El segundo sirve para analizar una variable politómica; por ejemplo, ¿ha cambiado en la última década la proporción de españoles con ideología política de derecha, de centro y de izquierda? Por tanto, para elegir entre los dos contrastes propuestos para analizar una variable categórica únicamente hay que considerar si la variable tiene dos categorías o más. Estudiaremos ambos contrastes en este volumen.

---

[6] No falta quien considera (ver, en el Capítulo 1, el apartado *Rol de las escalas de medida*) que este criterio de clasificación es inapropiado. Pero lo cierto es que la *naturaleza categórica o cuantitativa de las variables* condiciona el tipo de estadísticos que permiten obtener información útil de los datos. Con variables nominales como, por ejemplo, el *lugar de nacimiento* no tiene sentido calcular medias: ¿cuál es la media de *Andalucía*, *Aragón*, *Asturias*, ..., *Valencia*? Y con variables cuantitativas como, por ejemplo, la *edad* no tiene mucha utilidad preguntarse qué porcentaje de sujetos tiene una determinada edad (si la variable está medida con suficiente precisión, no habrá repeticiones o habrá muy pocas), es más útil conocer el centro, la dispersión y la forma de la distribución. Por tanto, los estadísticos que ofrecen información útil con variables categóricas no son los mismos que los que ofrecen información útil con variables cuantitativas. La clasificación propuesta tiene en cuenta esta circunstancia incorporando la naturaleza de las variables como un criterio más.

Para analizar **una variable cuantitativa** se proponen cinco contrastes. Tres de ellos sirven para contrastar hipótesis sobre el *centro* de una distribución: ¿es cierto que el peso medio de los recién nacidos de madres fumadoras no alcanza los 2,5 kg? El cuarto, para contrastar hipótesis sobre la dispersión de una distribución: ¿ha cambiado la varianza del peso de los recién nacidos? Y el quinto, para contrastar hipótesis sobre la *forma de la distribución*: ¿puede afirmarse que el peso de los recién nacidos se distribuye normalmente? La elección entre los tres primeros depende del tipo de supuestos que puedan establecerse sobre la forma de la población muestreada. La prueba *T* asume que la población muestreada es normal, pero, según veremos, este supuesto puede pasarse por alto si el tamaño muestral es lo bastante grande. Cuando es necesario trabajar con muestras pequeñas y no puede asumirse que la población de partida es normal, la prueba de Wilcoxon es una excelente alternativa a la prueba *T*, pero exige que la distribución poblacional de la variable sea simétrica. Si tampoco puede asumirse población simétrica, puede recurrirse a la prueba de los signos. En este volumen estudiaremos la *prueba T de Student para una muestra*.

En lo relativo al estudio de dos variables, estudiaremos contrastes tanto para efectuar comparaciones como para estudiar relaciones. Para analizar **dos variables categóricas** se proponen varios contrastes. Para elegir entre ellos hay que prestar atención, en primer lugar, a si interesa contrastar la hipótesis de independencia o la de homogeneidad marginal. Supongamos que se pregunta a una serie de personas su opinión sobre la eutanasia y sobre el aborto (a favor, en contra). El contraste de la hipótesis de independencia (prueba X^2 de Pearson, 'odds ratio', etc.) permite estudiar si las personas que están a favor de la eutanasia también lo están a favor del aborto. El contraste de la hipótesis de homogeneidad marginal (prueba de McNemar) permite estudiar si la proporción de personas que están a favor de la eutanasia difiere o no de la proporción de personas que están a favor del aborto (la hipótesis de homogeneidad marginal sólo tiene sentido si las dos variables analizadas tienen las mismas categorías). Para estudiar, no si existen diferencias o relaciones, sino el tamaño de las diferencias o la intensidad de las relaciones, propondremos diferentes medidas de asociación. Y también prestaremos atención al acuerdo como un caso especial de la asociación entre variables. En este volumen estudiaremos *la prueba X^2 de Pearson sobre independencia o igualdad de proporciones*.

Para analizar **una variable categórica y una cuantitativa** se proponen cuatro contrastes: dos para cuando la variable categórica tiene sólo dos niveles (hombres, mujeres; grupo experimental, grupo control; etc.) y otros dos para cuando la variable categórica tiene más de dos niveles (nivel de estudios –primarios, secundarios, medios, superiores –; tipo de tratamiento –farmacológico, mixto, control –; etc.). La elección entre cada par de contrastes depende de si puede asumirse o no que las distribuciones poblacionales son normales y del tamaño muestral (si puede asumirse normalidad o el tamaño muestral es grande se utiliza la prueba *T* de Student y el análisis de varianza; en caso contrario, la prueba de Mann-Whitney y la prueba de Kruskal-Wallis). En este volumen estudiaremos la *prueba T de Student para muestras independientes*.

Por último, para analizar dos **variables cuantitativas** se proponen varios contrastes. Tres de ellos sirven para *compararlas*: ¿ha disminuido el nivel de depresión de los pacientes tras aplicar un determinado tratamiento? Los demás sirven para *relacionarlas*: ¿tiene algo que ver la inteligencia con el rendimiento? En ambos casos, la elección entre los diferentes contrastes depende de si puede o no asumirse que las distribuciones poblacionales de ambas variables son normales y del nivel de medida de las variables. En este volumen estudiaremos *la prueba T* y el *coeficiente de correlación de Pearson*.

Cuadro 8.1. Clasificación de los contrastes de hipótesis para una y dos variables (los contrastes que aparecen en cursiva se estudian en este volumen; el resto se estudia en el siguiente volumen)

Una variable categórica:

- Si la variable es dicotómica:
  Prueba binomial o contraste sobre una proporción.
- Si la variable es politómica:
  Prueba X^2 de Pearson sobre bondad de ajuste.

Una variable cuantitativa:

- Para estudiar el centro de la distribución:
  Prueba T de Student para una muestra.
  Prueba de Wilcoxon para una muestra.
  Prueba de los signos para una muestra.
- Para estudiar la dispersión de la distribución:
  Contraste sobre una varianza.
- Para estudiar la forma de la distribución:
  Prueba de Kolmogorov-Smirnov sobre bondad de ajuste.

Dos variables categóricas:

- Para contrastar la hipótesis de independencia:
  Prueba X^2 de Pearson sobre independencia o igualdad de proporciones.
  Índices de riesgo y 'odds ratio'.
- Para contrastar la hipótesis de homogeneidad marginal (tablas 2×2):
  Prueba de McNemar.
- Para cuantificar el grado de asociación:
  Medidas de asociación (variables nominales, variables ordinales, acuerdo).

Una variable categórica y una cuantitativa:

- Si la variable categórica tiene 2 niveles:
  Prueba T para muestras independientes.
  Prueba *U* de Mann-Whitney.
- Si la variable categórica tiene más de dos niveles:
  Análisis de varianza de un factor.
  Prueba *H* de Kruskal-Wallis.

Dos variables cuantitativas:

- Para compararlas:
  Prueba T de Student para muestras relacionadas.
  Prueba de Wilcoxon para dos muestras.
  Prueba de los signos para dos muestras.
- Para relacionarlas:
  Coeficiente de correlación R_{XY} de Pearson.
  Coeficientes de correlación para variables ordinales.

Apéndice 8

Consideraciones sobre el nivel crítico (valor *p*)

Hemos definido el nivel de significación α como la probabilidad asociada al conjunto de valores considerados incompatibles con H_0. En ese sentido, podemos entender el nivel de significación α como el riesgo máximo que estamos dispuestos a asumir al tomar la decisión de rechazar una hipótesis nula que en realidad es verdadera. Esa probabilidad α suele establecerse *antes* de efectuar el contraste para evitar que influya en la decisión final. Realizar un contraste estableciendo previamente un nivel de significación α es lo que se viene haciendo hace décadas en la mayor parte de las áreas de conocimiento por la mayor parte de los investigadores. Sin embargo, esto no significa que esta forma de proceder esté libre de problemas. Los tiene, y no pequeños. Dos de ellos son éstos:

1. La decisión sobre H_0 puede depender decisivamente del nivel de significación establecido. Es posible mantener una hipótesis con $\alpha = 0{,}01$ y, sin embargo, rechazarla con $\alpha = 0{,}05$.
2. Decidir si H_0 es o no falsa no ofrece información sobre el grado en el que la evidencia empírica se muestra incompatible con esa hipótesis.

En relación con el primero de estos problemas, aunque es cierto que existe un acuerdo generalizado acerca de que α debe ser un valor pequeño, *cómo de pequeño* es algo que nos vemos obligados a establecer de forma arbitraria. Y aunque los niveles de significación habitualmente utilizados son 0,05 y 0,01, no existe ningún argumento serio que impida utilizar otro nivel de significación como 0,03 o 0,005. En principio, si se considera que el error consistente en rechazar una hipótesis verdadera tiene consecuencias *graves*, se debe adoptar para α un valor más pequeño que si se considera que las consecuencias no son graves. Pero ocurre que, al hacer más pequeño el valor de α, disminuye la probabilidad de rechazar una H_0 falsa; lo cual significa que, para evitar un tipo de error (rechazar una hipótesis verdadera) se está aplicando un remedio que aumenta la probabilidad de cometer otro tipo de error (no rechazar una hipótesis falsa).

Podríamos servirnos de conocimientos previos (resultados informados por otras investigaciones o por trabajos piloto; predicciones derivadas de alguna teoría; etc.) para establecer un nivel de significación más grande o más pequeño dependiendo de si esos conocimientos previos apuntan en la dirección de H_0 o en otra dirección. Pero incluso así, el valor que se adopte para α seguiría siendo arbitrario (al menos, en un rango de posibles valores asumibles con cierta coherencia).

Y siendo α un valor arbitrariamente establecido, resulta obligado hacer referencia al primero de los problemas mencionados. Aclaremos esto con un ejemplo. Supongamos que, para contrastar la hipótesis nula H_0: $\mu = 10$ frente a la alternativa H_1: $\mu \neq 10$, se adopta un nivel de confianza de 0,95 (es decir, $\alpha = 0{,}05$). Supongamos que, con ese nivel de confianza, la zona crítica está formada por los valores mayores que 1,96 y los menores que $-1{,}96$. Supongamos finalmente que el estadístico del contraste vale 2,14. Puesto que este valor cae en la zona crítica, decidimos rechazar H_0. Lo curioso de este contraste es que, si en lugar de adoptar para α un valor de 0,05 se adopta un valor de 0,01, la zona crítica estará formada por los valores mayores que 2,58 y los menores que $-2{,}58$; lo cual significa que el valor del estadístico del contraste (2,14) no caerá en la zona crítica y eso llevará a mantener H_0. En consecuencia, con $\alpha = 0{,}05$ tomaremos la decisión de rechazar H_0; y con $\alpha = 0{,}01$ tomaremos la decisión de mantener H_0. Esto ocurre porque la probabilidad de encontrar valores iguales o mayores que el encontrado vale $P(Z \geq 2{,}14) = 0{,}0162$, y ese valor se encuentra entre 0,025 y 0,005 ($\alpha/2$). Hace falta un nivel de significación mínimo de 0,0324 (= $0{,}0162 + 0{,}0162$, pues el contraste es bilateral) para que el valor del estadístico del contraste lleve al rechazo de H_0. Cualquier valor α mayor que 0,0324 llevará a tomar la decisión de mantener H_0.

Ya nos hemos referido al nivel crítico (valor p) como un número que expresa el grado de compatibilidad entre la hipótesis nula (H_0) y los datos (D). Más concretamente, el nivel crítico es la *probabilidad asociada al estadístico de contraste cuando H_0 es verdadera*. Se trata, por tanto, de una probabilidad condicional (ver ecuación [1.1]). En términos generales, en un contraste *unilateral*, el nivel crítico es la probabilidad asociada a los valores *mayores* (contraste unilateral derecho) o *menores* (contraste unilateral izquierdo) que el del estadístico del contraste; en un *contraste bilateral*, el nivel crítico es la probabilidad asociada a los valores que se encuentran tan alejados de H_0 como, al menos, el valor del estadístico del contraste[7]. Por tanto, el nivel crítico p se obtiene, a diferencia del nivel de significación α, *después* de efectuar el contraste, es decir, una vez calculado el estadístico del contraste, y siempre bajo el supuesto de que H_0 es verdadera.

Algunos investigadores, en lugar de establecer a priori un nivel de significación α, prefieren esperar a obtener el nivel crítico para tomar una decisión sobre H_0: si el nivel crítico es pequeño, la decisión será rechazarla; si el nivel crítico es grande, la decisión será mantenerla. Por supuesto, de nuevo nos encontramos con la arbitrariedad de tener que determinar cuándo un nivel crítico es *grande* y cuándo es *pequeño*. Pero este problema tiene mejor salida que el de establecer a priori un valor para α. Una regla bastante razonable podría ser ésta: (1) rechazar H_0 si el nivel crítico es claramente menor que 0,05; (2) mantenerla si es claramente mayor que 0,05; (3) repetir el contraste con una muestra diferente si el nivel crítico toma un valor en torno a 0,05. Por supuesto, la importancia o gravead de cada tipo de error y el conocimiento previo que se tenga sobre la hipótesis contrastada podrían servir para matizar el significado de las expresiones *claramente mayor*, *claramente menor* y *en torno a* referidas en la regla propuesta.

La utilización del nivel crítico p en lugar del nivel de significación α tiene una ventaja adicional que permite superar, en parte, el segundo de los inconvenientes asociados a la utilización de un nivel de significación establecido a priori. El nivel crítico no sólo sirve para tomar una decisión sobre H_0, sino que su tamaño informa sobre el grado de compatibilidad o discrepancia existente entre esa hipótesis y la evidencia muestral disponible. Un nivel crítico de 0,70, por ejemplo, está indicando que el resultado muestral obtenido es perfectamente compatible con la hipótesis planteada; es decir, un nivel crítico de ese tamaño está indicando que, si asumimos que la H_0 planteada es verdadera, la probabilidad de encontrar un resultado muestral como el encontrado o más extremo vale 0,70. Un nivel crítico de 0,05 está indicando que el resultado muestral observado es *poco* compatible con H_0. Un nivel crítico de 0,000001 está indicando que el resultado muestral observado se encuentra tan alejado de la predicción efectuada en H_0 que, siendo H_0 verdadera, sólo hay una posibilidad entre un millón de encontrar un resultado semejante; con un nivel crítico de 0,000001 podríamos sentirnos razonablemente seguros de que la H_0 planteada es falsa. Puede decirse, por tanto, que el tamaño del nivel crítico está informando del grado en el que la evidencia empírica se muestra incompatible con la H_0 planteada (información ésta que se pasa por alto cuando la decisión se limita a mantener o rechazar H_0 a partir de un nivel de significación previamente establecido).

No obstante, no debe olvidarse que el tamaño del error típico de la distribución muestral de un estadístico depende directamente del tamaño de la muestra utilizada: cuanto mayor es el tamaño muestral, menor es el error típico de un estadístico (ver ecuaciones [6.1] y [6.7]). Por tanto, un mismo efecto (es decir, una misma diferencia, o una misma relación) tendrá asociado un estadístico tanto más extremo en su distribución muestral (es decir, tanto más alejado de la predicción formulada en H_0) cuanto mayor sea el tamaño muestral. En consecuencia, a medida que el tamaño muestral vaya aumentando, el nivel crítico irá tendiendo a 0; y esto podría llevar a pensar, erróneamente, que existe una discrepancia importante entre la hipótesis nula y los datos.

---

[7] En los contrastes en los que se utilizan las dos colas de la distribución muestral, el nivel crítico p se obtiene, generalmente, multiplicando por 2 la probabilidad asociada a los valores mayores (si el estadístico del contraste cae en la cola derecha) o menores (si el estadístico del contraste cae en la cola izquierda) que el valor concreto del estadístico del contraste. Pero existen contrastes bilaterales en los que la zona crítica está situada, toda ella, en la cola derecha de la distribución muestral. En estos casos, el nivel crítico es la probabilidad asociada a los valores mayores que el estadístico del contraste.

Todo esto sugiere que la utilización del nivel crítico como una medida del grado de discrepancia entre la hipótesis nula y los datos, aunque puede servir para mantener o rechazar una hipótesis particular, no sirve como una medida de la magnitud del efecto encontrado (el tamaño de una diferencia, la intensidad de una relación; etc.). Para esto último es necesario utilizar medidas que no dependan del tamaño muestral. Precisamente la búsqueda de una medida de ese tipo es lo que se ha venido haciendo en los últimos años bajo la denominación genérica de *medidas del tamaño del efecto*.

Decidir si una hipótesis es o no falsa no constituye, en muchas ocasiones, un criterio *suficiente* para determinar si el estudio realizado contribuye o no de forma significativa al desarrollo de una teoría o de una línea de investigación. Esto se debe a que la decisión a la que se llega en un contraste de hipótesis sobre la base del grado de discrepancia existente entre la hipótesis nula y los datos depende en buena medida, según acabamos de señalar, del tamaño de la muestra concreta utilizada. Tamaños muestrales grandes pueden llevar a considerar como estadísticamente significativas discrepancias insignificantes; y tamaños muestrales muy pequeños pueden llevar a considerar estadísticamente no significativas discrepancias teóricamente relevantes[8].

Desde hace décadas se viene insistiendo en la conveniencia de acompañar la decisión típica de un contraste de hipótesis (mantener o rechazar H_0) con alguna medida del grado de discrepancia existente entre esa H_0 y la evidencia muestral disponible. Acabamos de destacar la importancia de la información que ofrece una medida de este tipo, pero no hemos ofrecido ninguna solución aceptable (el nivel crítico como medida de esa discrepancia no es una buena solución). En el segundo volumen estudiaremos algunas de estas medidas. De momento, basta con que tomemos conciencia de la importancia de cuantificar el *tamaño del efecto* y de la conveniencia de utilizar esa cuantificación como un complemento de la información que ofrece un contraste de hipótesis.

Ejercicios

8.1. ¿Cuáles de las hipótesis que siguen están bien formuladas?

a. H_0: $\mu \leq 3$; H_1: $\mu \geq 3$.
b. H_0: $\mu > 3$; H_1: $\mu < 3$.
c. H_0: $\pi \geq 0{,}5$; H_1: $\pi < 0{,}5$.
d. H_0: $\pi \leq 0{,}5$; H_1: $\pi > 0{,}5$.
e. H_0: $\mu \geq 3$; H_1: $\mu \leq 3$.
f. H_0: $\pi \leq 0{,}5$; H_1: $\pi \neq 0{,}5$.
g. H_0: $\mu \neq 3$; H_1: $\mu = 3$.
h. H_0: $\mu < 3$; H_1: $\mu \geq 3$.

8.2. En un contraste de hipótesis, la probabilidad mínima con la cual se puede rechazar una hipótesis nula que es verdadera se denomina:

a. Nivel de significación.
b. Nivel crítico.
c. Nivel de confianza.
d. Nivel de riesgo.
e. Zona crítica.

[8] Decir que un resultado muestral es *estadísticamente significativo* no implica necesariamente que ese resultado es *relevante* desde un punto de vista sustantivo, ya sea teórico o aplicado. Volveremos sobre esta cuestión en el segundo volumen. En León (1984) puede encontrarse una interesante discusión sobre el significado del concepto *significativo* en el contexto del análisis de datos y fuera de él.

8.3. Supongamos que se desea evaluar la eficacia de un tratamiento. Para ello, se selecciona aleatoriamente una muestra de pacientes y se forman, también aleatoriamente, dos grupos: experimental y control. Al grupo experimental se le aplica el tratamiento; al grupo control se le aplica un placebo. Tras recoger los datos y comparar los grupos se obtiene un resultado significativo ($p = 0{,}001$). Teniendo en cuenta este escenario, señalar como verdadera o falsa cada una de las siguientes afirmaciones:

a. Se ha conseguido probar definitivamente la eficacia del tratamiento.
b. Se conoce o puede deducirse la probabilidad de que la hipótesis nula sea verdadera.
c. Se conoce o puede deducirse la probabilidad de que la hipótesis nula sea falsa.
d. Si se decide rechazar la hipótesis nula, se conoce la probabilidad de que la decisión sea incorrecta.
e. Si se repitiera el experimento un gran número de veces, sabemos que obtendremos un resultado significativo en el 99,9% de las veces.
f. Si se mantiene la hipótesis nula, puede concluirse que los grupos no difieren.

8.4. Un estadístico V tiene la función de probabilidad acumulada que muestra la tabla. Llevado a cabo un contraste unilateral izquierdo con una determinada muestra obtenemos para el estadístico V un valor de −1.

V	−1	−0,5	0	0,5	1	1,5	2
$F(V\|H_0)$	0,03	0,05	0,37	0,65	0,90	0,97	1

a. Establecer una regla de decisión en términos de probabilidad.
b. ¿Qué decisión debe tomarse sobre H_0? ¿Por qué?
c. ¿Cuánto vale el nivel crítico p?

8.5. El estadístico n_1 (número de aciertos) se distribuye según se indica en la siguiente tabla. Si se plantea, con $\alpha = 0{,}05$, un contraste unilateral derecho sobre H_0: $\pi_1 = 0{,}40$, ¿qué decisión debe tomarse sobre H_0 si en una muestra concreta se obtiene $n_1 = 4$?

n_1	0	1	2	3	4
$f(n_1 \mid \pi = 0{,}40)$	0,130	0,345	0,345	0,154	0,026

8.6. Un investigador afirma que, entre los estudiantes universitarios, ellas fuman más que ellos. Tras efectuar una encuesta, ha comparado la proporción de fumadoras con la proporción de fumadores (H_0: $\pi_{ellas} \leq \pi_{ellos}$; H_1: $\pi_{ellas} > \pi_{ellos}$) y ha obtenido, para el estadístico del contraste, un valor $T = 2{,}681$. La siguiente tabla ofrece la función de distribución (probabilidades acumuladas) de algunos de los valores del estadístico T:

T	−0,539	0,000	0,539	1,356	1,782	2,179	2,681	3,055
$F(T)\|H_0$	0,300	0,500	0,700	0,900	0,950	0,975	0,990	0,995

Con $\alpha = 0{,}05$:

a. ¿Puede afirmarse que los datos confirman la hipótesis del investigador? ¿Por qué?
b. ¿A partir de qué nivel de significación podría rechazarse H_0?

8.7. Un terapeuta afirma que una determinada terapia consigue recuperaciones aceptables en más del 80% de los pacientes tratados. Un colega suyo piensa que la proporción de recuperaciones aceptables es menor que el 80%. Ambos realizan un estudio para contrastar sus respectivas hipótesis con $\alpha = 0{,}05$:

a. ¿Qué hipótesis estadísticas debe plantear cada terapeuta?
b. Al contrastar su hipótesis nula el primer terapeuta obtiene un nivel crítico $p = 0{,}818$. Utilizando un nivel de confianza de 0,95, ¿qué decisión debe tomar? ¿Por qué?
c. Al contrastar su hipótesis nula el segundo terapeuta obtiene un nivel crítico $p = 0{,}002$. Utilizando un nivel de confianza de 0,95, ¿qué decisión debe tomar? ¿Por qué?
d. ¿Cuál de los dos terapeutas tiene razón?, ¿tienen razón los dos?, ¿no tiene razón ninguno de los dos?

8.8. En 1990 fumaba el 30% de los universitarios madrileños. Un investigador cree que en los últimos años ese porcentaje ha aumentado. Para comprobarlo, selecciona una muestra aleatoria de universitarios madrileños y obtiene un estadístico al que, asumiendo H_0: $\pi_F = 0{,}30$ verdadera, corresponde el *centil* 93.

a. Plantea las hipótesis estadísticas del contraste.
b. ¿Qué decisión debe tomarse sobre H_0 con $\alpha = 0{,}05$? ¿Por qué?

8.9. Supongamos que se contrasta H_0: $\mu_Y \geq 0$ frente a H_1: $\mu_Y < 0$ y, en una muestra aleatoria simple, se obtiene un estadístico de contraste $T = -2$. Sabiendo que $P(T < -2) = 0{,}005$ y utilizando un nivel de significación $\alpha = 0{,}01$, ¿qué decisión debe tomarse sobre H_0 y por qué?

a. Rechazarla porque $-2 < 0$.
b. Mantenerla porque $0{,}01 < 0{,}995$.
c. Mantenerla porque $-2 < 0{,}01$.
d. Rechazarla porque $0{,}005 < 0{,}01$.
c. Mantenerla porque $P(T < -2) > \alpha$.

8.10. Al comparar las medias de dos grupos mediante un contraste unilateral derecho el estadístico del contraste ha tomado el valor $T = 2{,}63$. Sabiendo que $P(T > 2{,}63) = 0{,}075$ y utilizando un nivel de significación de 0,05:

a. Se debe rechazar H_0 porque T cae en la zona crítica.
b. Se debe mantener H_0 porque $0{,}075 > 0{,}05$.
c. Se debe rechazar H_0 porque $0{,}075 > 0{,}05$.
d. Se debe concluir que las medias poblacionales difieren entre sí.
e. Se debe concluir que las medias muestrales son iguales.

8.11. En un contraste de hipótesis unilateral derecho se obtiene, para el estadístico del contraste, un valor $H = 6{,}13$. Sabiendo que $P(H < 6{,}13) = 0{,}05$:

a. La decisión razonable es mantener H_0.
b. La decisión razonable es rechazar H_0.
c. La probabilidad de rechazar H_0 siendo verdadera vale 0,05.
d. Se puede rechazar H_0 con una probabilidad de equivocarse de 0,05.
e. Al mantener H_0 siendo verdadera, la probabilidad de equivocarse vale al menos 0,05.

8.12. Para contrastar una determinada hipótesis nula se han utilizado dos estadísticos: V y W. Se sabe que V se distribuye según el modelo de probabilidad t de Student y que W se distribuye según el modelo de probabilidad normal $N(0, 1)$. En una muestra aleatoria concreta se ha obtenido $V = W = k$. Según esto, siendo k un valor cualquiera y dado un mismo nivel de significación (señalar la/s alternativa/s correcta/s):

a. Si se mantiene H_0 con V, es imposible rechazarla con W.
b. Si se rechaza H_0 con V, necesariamente se rechazará con W.
c. Es más probable rechazar H_0 con V que con W.
d. Si se mantiene H_0 con V, necesariamente se mantendrá con W.
e. Si se rechaza H_0 con V, es posible mantenerla con W.

Soluciones

8.1. Sólo son correctas las alternativas *c* y *d*. Las alternativas *a*, *e*, *g* y *h* son incorrectas porque las hipótesis alternativas contienen el signo "=". En la alternativa *b* falta el signo "=" en la hipótesis nula. Y las hipótesis de la alternativa *f* no son exclusivas.

8.2. La probabilidad mínima con la que puede ser rechazada una hipótesis verdadera es justamente la probabilidad asociada al estadístico de contraste obtenido. Y esa probabilidad no es otra cosa que el nivel crítico (valor p). La alternativa correcta es la *b*.

8.3. Todas las afirmaciones son falsas. Un contraste de hipótesis se basa en probabilidades y, por tanto, no es una prueba definitiva de nada (esto descarta la alternativa *a*).

El nivel crítico es una probabilidad condicional referida a los datos (al estadístico del contraste). Esto no debe confundirse con una probabilidad referida a las hipótesis (esto descarta las alternativas *b* y *c*).

La alternativa *d* también es una afirmación referida a la probabilidad de una hipótesis, pues una afirmación sobre la probabilidad de tomar una decisión errónea cuando se decide rechazar una hipótesis es una afirmación sobre la probabilidad de que la hipótesis rechazada sea verdadera (esto descarta la alternativa *d*).

La alternativa *e* sólo tendría sentido si la hipótesis nula fuera verdadera, pero eso es algo que se desconoce en un contraste de hipótesis. El nivel crítico es una probabilidad condicional; es decir, no es una probabilidad referida a los datos en sí mismos, sino una probabilidad referida a los datos cuando la hipótesis nula es verdadera; en la alternativa *e* se está confundiendo $P(D)$ con $P(D|H_0)$, pues se está afirmando $1 - P(D)$ a partir de $P(D|H_0)$.

Por último, la alternativa *f* incluye una afirmación de la hipótesis nula que no puede hacerse en un contraste de hipótesis (recuérdese la falacia de la afirmación del consecuente).

8.4. *a*. Rechazar H_0 si $P(V \geq V_k) \leq 0{,}05$.
b. Rechazarla. Porque $P(V \leq -1) < 0{,}05$.
c. $p = P(V \leq 1) = 0{,}03$.

8.5. Rechazarla, pues $P(n_1 \geq 4) = 0{,}026 < 0{,}05$.

8.6. *a*. Puesto que el nivel crítico ($p = P(T > 2{,}681) = 1 - 0{,}99 = 0{,}01$) es menor que el nivel de significación (0,05), la decisión razonable es rechazar H_0. Esto significa que puede concluirse que *ellas* fuman más que *ellos*, lo cual confirma la hipótesis del investigador.
b. $p = 1 - 0{,}99 = 0{,}01$.

8.7. *a*. El primer terapeuta: H_0: $\pi_R \leq 0{,}80$; H_1: $\pi_R > 0{,}80$ (R = recuperación).
El segundo psicólogo: H_0: $\pi_R \geq 0{,}80$; H_1: $\pi_R < 0{,}80$.
b. Mantener H_0, porque $p = 0{,}818 > \alpha = 0{,}05$.
c. Rechazar H_0, porque $p = 0{,}002 < \alpha = 0{,}05$.
d. Puesto que sólo el segundo terapeuta rechaza su hipótesis nula, puede concluirse que la proporción de recuperaciones es menor que 0,80. Por tanto, tiene razón el segundo terapeuta.

8.8. *a*. H_0: $\pi_F \leq 0{,}30$; H_1: $\pi_F > 0{,}30$.
b. Mantenerla, porque $p = 0{,}07 > 0{,}05$.

8.9. La decisión correcta es rechazar H_0 porque $p = 0{,}005 < \alpha = 0{,}01$. Por tanto, la única alternativa correcta es la *d*.

8.10. Puesto que el contraste es unilateral derecho, la zona crítica está, toda ella, situada en la cola derecha de la distribución muestral. El estadístico del contraste toma un valor tal que por encima de él queda un área de tamaño 0,075, es decir, de tamaño superior al área correspondiente a la zona crítica (que vale $\alpha = 0{,}05$). En consecuencia, el estadístico del contraste cae fuera de la zona crítica y eso debe llevarnos a tomar la decisión de mantener H_0. Esto descarta las alternativas *a*, *c* y *d*.

La alternativa *e* también incluye una afirmación falsa, pues las inferencias estadísticas nunca se efectúan sobre valores muestrales, sino sobre valores poblacionales (la alternativa *e* sería falsa incluso aunque la afirmación que se hace estuviera referida a las medias poblacionales, pues se estaría afirmando una hipótesis nula y, con ello, se estaría cayendo en la falacia derivada de afirmar el consecuente).

Sólo queda la alternativa *b*, que es la correcta: no puede rechazarse H_0. Y no puede rechazarse porque, según se afirma en la alternativa *b*, el nivel crítico $p = 0{,}075$ es mayor que el nivel de significación establecido ($\alpha = 0{,}05$).

8.11. Puesto que $P(H < 6{,}13) = 0{,}05$, el estadístico del contraste, H, se encuentra en la cola izquierda de su distribución muestral (tiene asociado, por tanto, un nivel crítico de 0,95). Es decir, el estadístico H se encuentra justamente en el lado opuesto al de la zona de rechazo. Esto significa que la decisión apropiada es mantener la hipótesis nula: el nivel crítico asociado al estadístico de contraste, 0,95, es mucho mayor que cualquier nivel de significación razonable que se pueda establecer. Esto lleva a considerar como correcta la alternativa *a*, al tiempo que permite descartar las alternativas *b* y *d*.

Un nivel crítico (valor p) no permite conocer la probabilidad asociada a las hipótesis de un contraste. Esto descarta las alternativas *c* y *e*. Por otro lado, en relación con la última alternativa hay que decir, entre otras cosas, que mantener una hipótesis verdadera no es un error, por lo que no se puede hablar de probabilidad de equivocarse.

8.12. La distribución t de Student (ver Apéndice 5) acumula en sus colas más casos que la distribución normal. Esto significa que, para un mismo nivel de significación α, el punto crítico de una distribución t con un número finito de grados de libertad estará más alejado del centro de la distribución que el punto crítico equivalente en la distribución normal estandarizada. En consecuencia: (1) siempre que se rechace H_0 con V se rechazará con W; (2) siempre que se mantenga H_0 con W se mantendrá con V; y (3) es posible mantener con V y rechazar con W. Por tanto, es más probable rechazar con W que con V. Este razonamiento nos deja como única alternativa correcta la *b*.

9

Inferencia con una variable

En los Capítulos 3 al 5 hemos estudiado qué puede hacerse con una variable desde el punto de vista *descriptivo*. Ahora que ya conocemos los fundamentos de la inferencia estadística ha llegado el momento de estudiar qué puede hacerse desde el punto de vista *inferencial*.

La distinción que hemos venido haciendo entre variables *categóricas* y *cuantitativas* no es superficial o caprichosa. Ya hemos tenido ocasión de comprobar que las herramientas descriptivas que permiten obtener información útil de unos datos son unas u otras dependiendo de que la variable analizada sea categórica o cuantitativa; y también hemos tenido ocasión de comprobar que las distribuciones teóricas de probabilidad, incluidas las distribuciones muestrales, poseen peculiaridades propias dependiendo de que la variable original sea categórica o cuantitativa. Esta distinción básica entre variables categóricas y cuantitativas sigue siendo válida también a la hora de clasificar y elegir las herramientas estadísticas concretas que permitirán efectuar inferencias útiles.

También hemos venido insistiendo en la idea de que describir correctamente una variable exige prestar atención a tres características de su distribución: *centro*, *dispersión* y *forma*. Prestar atención a estas tres características sigue siendo útil también desde el punto de vista inferencial, pues, salvo muy contadas excepciones, todas las herramientas inferenciales diseñadas para analizar una sola variable se centran en alguna de esas tres características.

Con **una variable categórica** suele interesar estudiar cómo se reparten las frecuencias entre las diferentes categorías de la variable. Si la variable es dicotómica o dicotomizada (sólo dos categorías), esto significa dirigir la atención al *número* o *proporción de éxitos*; y, dado que una proporción es una media, esto equivale a estudiar el *centro* de la distribución. Si la variable es politómica (más de dos categorías), estudiar cómo se reparten las frecuencias entre las diferentes categorías de la variable significa estudiar la *forma* de la distribución.

Con **una variable cuantitativa** suele interesar estudiar tanto el centro de su distribución (a través de algún estadístico de tendencia central), como su grado de dispersión (a través de algún estadístico de dispersión) y la forma de su distribución (generalmente para valorar si se parece o no a una distribución normal o a alguna otra distribución de probabilidad teórica conocida).

Estas consideraciones nos ponen en la pista del tipo de herramientas que utilizaremos para hacer inferencias con una sola variable. Con una variable categórica dicotómica haremos inferencias sobre el parámetro *proporción*; con una variable categórica politómica, sobre la

forma de su distribución; y con una variable cuantitativa, sobre el centro, sobre la dispersión y sobre la forma de su distribución.

Los procedimientos para realizar estas inferencias están clasificados en el Cuadro 8.1 del capítulo anterior. Este capítulo se centra en tres de ellos: el *contraste sobre una proporción* (para una variable dicotómica), el *contraste sobre bondad de ajuste* (para una variable politómica) y el *contraste sobre una media* (para una variable cuantitativa). El resto de procedimientos disponibles para una variable se estudian en el segundo volumen.

El contraste sobre una proporción (prueba binomial)

En las ciencias sociales y de la salud es frecuente encontrarse con variables dicotómicas (variables categóricas que toman sólo dos valores): acierto-error, verdadero-falso, recuperados-no recuperados, a favor-en contra, etc. Muchas de estas variables se obtienen registrando la *presencia-ausencia* de algún evento (por ejemplo, la presencia-ausencia de un tratamiento o de un síntoma); otras muchas se crean dicotomizando variables cuantitativas (por ejemplo, aprobados-suspensos, hipertensos-no hipertensos, etc.). Con este tipo de variables es costumbre codificar con "unos" la presencia de la característica, los aciertos, los recuperados, etc., y con "ceros" la ausencia de la característica, los errores, los no recuperados, etc. Y, aunque es práctica bastante generalizada llamar *éxito* a uno de estos resultados y *fracaso* al otro, en muchos contextos es preferible utilizar una terminología que capte con mayor precisión el significado del ensayo, como presencia-ausencia, o sucede-no sucede.

Ya hemos señalado (Capítulo 6, apartado *Distribución muestral del estadístico proporción*) que, al trabajar con este tipo de variables, lo habitual en centrarse en el *número* o *proporción de éxitos* (o de fracasos). También hemos estudiado ya la distribución muestral de los estadísticos n_1 = «número de éxitos» y P_1 = «proporción de éxitos», y hemos tenido ocasión de comprobar que ambos estadísticos se distribuyen según el modelo de probabilidad binomial con parámetros n (número de ensayos) y π_1 (proporción de éxitos). El modelo binomial, en consecuencia, ofrece las probabilidades asociadas a los estadísticos n_1 y P_1, y eso es todo lo que hace falta para poder diseñar un contraste de hipótesis basado en estos estadísticos.

Pero también hemos estudiado (ver Capítulo 6) que, a medida que n va aumentando, las distribuciones de n_1 y P_1 se aproximan a la normal con parámetros

$$\begin{aligned} E(n_1) &= \mu_{n_1} = n\pi_1, \quad & V(n_1) &= \sigma^2_{n_1} = n\pi_1(1-\pi_1), \quad & \sigma_{n_1} &= \sqrt{n\pi_1(1-\pi_1)} \\ E(P_1) &= \mu_{P_1} = \pi_1, \quad & V(P_1) &= \sigma^2_{P_1} = \pi_1(1-\pi_1)/n, \quad & \sigma_{P_1} &= \sqrt{\pi_1(1-\pi_1)/n} \end{aligned} \qquad [9.1]$$

por lo que, conforme n va aumentando, la transformación

$$Z = \frac{n_1 - n\pi_1}{\sqrt{n\pi_1(1-\pi_1)}} = \frac{P_1 - \pi_1}{\sqrt{\pi_1(1-\pi_1)/n}} \qquad [9.2]$$

tiende a distribuirse $N(0, 1)$. Esto significa que también es posible utilizar la distribución normal para diseñar contrastes basados en los estadísticos n_1 y P_1. El Cuadro 9.1 ofrece, siguiendo la lógica general ya expuesta en el capítulo anterior, un resumen del contraste sobre una proporción (también conocido como *prueba binomial*). El resumen también incluye el intervalo de confianza para el parámetro π_1 y la forma de calcular el nivel crítico (valor p).

Este contraste sirve para tomar decisiones sobre el parámetro *proporción*[1]. Esto, traducido a la práctica, significa responder a preguntas del tipo: ¿consigue un nuevo tratamiento más recuperaciones de las que se vienen obteniendo con el tratamiento convencional?, ¿es posible afirmar que la proporción de sujetos que tienen intención de votar en las próximas elecciones es mayor que la registrada en las elecciones pasadas?, ¿ha acertado un sujeto más preguntas de las esperables por azar en una determinada prueba de rendimiento?, etc. El intervalo de confianza sirve para dar respuestas a estas mismas preguntas y, además, para obtener una estimación del valor aproximado del parámetro π_1.

Cuadro 9.1. Resumen del contraste de hipótesis sobre una proporción (prueba binomial)

1. *Hipótesis*:
   a. Contraste bilateral: H_0: $\pi_1 = k_0$; H_1: $\pi_1 \neq k_0$.
   b. Contraste unilateral derecho: H_0: $\pi_1 \leq k_0$; H_1: $\pi_1 > k_0$.
   c. Contraste unilateral izquierdo: H_0: $\pi_1 \geq k_0$; H_1: $\pi_1 < k_0$.
   (k_0 se refiere a la proporción concreta que interesa contrastar; lógicamente, al tratarse de una proporción, siempre tomará un valor comprendido entre 0 y 1).
2. *Supuestos*: la variable estudiada es dicotómica (o dicotomizada) y π_1 es la proporción de éxitos en la población (*éxito* hace referencia a una cualquiera de las dos categorías de la variable). De esa población se extrae una muestra aleatoria de n observaciones con probabilidad de éxito constante en cada extracción.
3. *Estadísticos del contraste*[2] (ver ecuaciones [9.1] y [9.2]):
   3.1. n_1 = «número de éxitos en los n ensayos».
   $P_1 = n_1/n$ = «proporción de éxitos en los n ensayos».
   3.2. $$Z = \frac{n_1 - n\pi_1}{\sqrt{n\pi_1(1-\pi_1)}} = \frac{P_1 - \pi_1}{\sqrt{\pi_1(1-\pi_1)/n}}$$
4. *Distribuciones muestrales*:
   4.1. n_1 y P_1 se distribuyen según el modelo binomial con parámetros n y π_1.
   4.2. Z se aproxima a la distribución $N(0, 1)$ a medida que n va aumentando[3].

---

[1] John Arbuthnott fue, al parecer, el primero en utilizar este procedimiento en 1710 para intentar demostrar que la proporción de varones nacidos en Londres en un determinado periodo de tiempo era significativamente mayor que la proporción de mujeres nacidas en ese mismo periodo.

[2] Tenemos tres estadísticos. Dos de ellos (n_1 y P_1) son, en realidad, el mismo y poseen una distribución muestral exacta (la binomial). El otro (Z) posee una distribución muestral aproximada (la normal tipificada). Los dos primeros son preferibles con muestras pequeñas (por ejemplo, con $n \leq 20$, que es el tope de la tabla binomial del Apéndice final). Z sólo debe utilizarse con muestras grandes (ver, en el Capítulo 5, el apartado *Aproximación de la distribución binomial a la normal*).

[3] Recordemos que, si n no es muy grande, la aproximación es un poco más exacta aplicando una pequeña modificación llamada *corrección por continuidad*, que consiste en sumar (si n_1 es menor que $n\pi_1$) o restar (si n_1 es mayor que $n\pi_1$) 0,5 a n_1 (o 0,5/n a P_1) para hacer el contraste algo más conservador:

$$Z = \frac{(n_1 \pm 0{,}5) - n\pi_1}{\sqrt{n\pi_1(1-\pi_1)}} = \frac{(P_1 \pm 0{,}5/n) - \pi_1}{\sqrt{\pi_1(1-\pi_1)/n}} \qquad [9.3]$$

5. *Reglas de decisión*:
   a. Contraste bilateral:
      a.1. Se rechaza H_0 si n_1 o P_1 toman un valor tan alejado de su valor esperado bajo H_0 que la probabilidad de obtener un valor tan alejado como ése o más es menor que $\alpha/2$.
      a.2. Se rechaza H_0 si $Z \leq Z_{\alpha/2}$ o $Z \geq Z_{1-\alpha/2}$.
   b. Contraste unilateral derecho:
      b.1. Se rechaza H_0 si n_1 o P_1 toman un valor tan grande que la probabilidad de obtener un valor como ése o mayor es menor que α.
      b.2. Se rechaza H_0 si $Z \geq Z_{1-\alpha}$.
   c. Contraste unilateral izquierdo:
      c.1. Se rechaza H_0 si n_1 o P_1 toman un valor tan pequeño que la probabilidad de obtener un valor como ése o más pequeño es menor que α.
      c.2. Se rechaza H_0 si $Z \leq Z_{\alpha}$.
6. *Nivel crítico* (valor *p*):
   a. Contraste bilateral:
      a.1. Con los estadísticos n_1 y P_1, el nivel crítico es el doble de la probabilidad de obtener un valor n_1 o P_1 tan alejado de su valor esperado bajo H_0 como el obtenido.
      a.2. Con el estadístico *Z*, $p = 2[P(Z \geq |Z_h|)]$, siendo Z_h el valor concreto que toma el estadístico *Z*.
   b. Contraste unilateral derecho:
      b.1. Con los estadísticos n_1 y P_1, el nivel crítico es la probabilidad de obtener un valor n_1 o P_1 tan grande como el obtenido o más grande.
      b.2. Con el estadístico *Z*, $p = P(Z \geq Z_h)$.
   c. Contraste unilateral izquierdo:
      c.1. Con los estadísticos n_1 y P_1, el nivel crítico es la probabilidad de obtener un valor n_1 o P_1 tan pequeño como el obtenido o más pequeño.
      c.2. Con el estadístico *Z*, $p = P(Z \leq Z_h)$.
7. *Intervalo de confianza* (ver [7.16] en Capítulo 7): $IC_{\pi_1} = P_1 \pm |Z_{\alpha/2}| \sqrt{P_1(1-P_1)/n}$.

Ejemplo. El contraste sobre una proporción

Al parecer, la sintomatología del 30% de los pacientes neuróticos remite espontáneamente durante los tres primeros meses del trastorno. Según esto, es lógico pensar que una terapia eficaz con este tipo de trastornos deberá conseguir a lo largo de los tres primeros meses un porcentaje de recuperaciones mayor que el que se produce de forma espontánea. Los resultados obtenidos con 20 sujetos a los que se les ha aplicado una determinada terapia indican que,

en los tres primeros meses, ha habido 9 recuperaciones. ¿Puede afirmase que el número de recuperaciones que se obtienen con esa terapia es mayor que el esperable por simple recuperación espontánea? ($\alpha = 0{,}05$).

Tenemos una variable dicotómica (pacientes recuperados, pacientes no recuperados) y una muestra de $n = 20$ observaciones de esa variable. Llamaremos π_1 a la proporción poblacional de la categoría *pacientes recuperados*. Hemos observado $n_1 = 9$ recuperaciones y, por tanto, la proporción observada de recuperaciones vale $P_1 = 9/20 = 0{,}45$.

Vamos a realizar un contraste sobre el parámetro π_1 para decidir si la verdadera proporción de pacientes recuperados con la terapia es o no superior a la que cabe esperar por simple recuperación espontánea (es decir, superior a 0,30):

1. *Hipótesis*: $H_0: \pi_1 \leq 0{,}30$; $H_1: \pi_1 > 0{,}30$.

   La hipótesis nula afirma que la terapia no es eficaz (proporción de recuperaciones igual o menor que 0,30); la hipótesis alternativa afirma que la terapia sí es eficaz (proporción de recuperaciones mayor que 0,30). Por tanto, se está planteando un contraste unilateral derecho: los valores incompatibles con H_0 son sólo los valores mayores que 0,30, no los menores. Debe tenerse en cuenta que la pregunta que se plantea en el enunciado del ejercicio hace referencia explícita a los valores *mayores* que 0,30 (¿... el número de recuperaciones... con esa terapia es mayor que el esperable...?).

2. *Supuestos*: muestra aleatoria de 20 observaciones con probabilidad constante 0,30 de que una observación cualquiera pertenezca a la categoría de *pacientes recuperados*.

3. *Estadísticos del contraste* (en un contraste concreto sólo es necesario utilizar uno de los tres estadísticos propuestos; aquí vamos a utilizar los tres con el único objetivo de ilustrar cómo trabajar con ellos):

   3.1. $n_1 = 9$.
   $P_1 = 0{,}45$.

   3.2. $$Z = \frac{9 - 20(0{,}30)}{\sqrt{20(0{,}30)(1 - 0{,}30)}} = \frac{0{,}45 - 0{,}30}{\sqrt{0{,}30(1 - 0{,}30)/20}} = 1{,}46.$$

4. *Distribuciones muestrales*:

   4.1. n_1 y P_1 se distribuyen binomialmente con parámetros $n = 20$ y $\pi_1 = 0{,}30$.

   4.2. Z se aproxima a $N(0, 1)$.

5. *Decisión*:

   5.1. Se rechaza H_0 si la probabilidad de obtener valores $n_1 \geq 9$ o $P_1 \geq 0{,}45$ es menor que $\alpha = 0{,}05$. Es decir, se rechaza H_0 si se verifica $P(n_1 \geq 9) < 0{,}05$; o, de forma equivalente, $P(P_1 \geq 0{,}45) < 0{,}05$. Consultando la tabla de la distribución binomial (ver Tabla B del Apéndice final), con $n = 20$ y $\pi_1 = 0{,}30$, se obtiene $P(n_1 \geq 9) = P(P_1 \geq 0{,}45) = 1 - F(8) = 1 - 0{,}887 = 0{,}113$ (no olvidar que la tabla ofrece probabilidades acumuladas).

   5.2. Se rechaza H_0 si $Z \geq Z_{0,95} = 1{,}645$.

   Estos resultados conducen a la siguiente decisión:

   - Como $P(n_1 \geq 9) = P(P_1 \geq 0{,}45) = 0{,}113$ es mayor que $\alpha = 0{,}05$, se mantiene H_0.
   - Como $Z = 1{,}46$ es menor que $Z_{0,95} = 1{,}645$, se mantiene H_0.

Es decir, tanto los estadísticos n_1 y P_1 como el estadístico Z llevan a la misma decisión. La conclusión razonable es que los datos disponibles no permiten afirmar que la proporción de mejoras que se obtiene con la terapia sea más alta que la proporción de mejoras que se producen de forma espontánea.

6. *Nivel crítico*[4] (valor p):

   6.1. Con los estadísticos n_1 y P_1 ya lo hemos calculado: $P(X \geq 9) = P(P \geq 0{,}45) = 0{,}113$.

   6.2. Con el estadístico Z: $p = P(Z \geq Z_h) = P(Z \geq 1{,}46) = 1 - F(1{,}46) = 1 - 0{,}9279 = 0{,}0721$.

7. *Intervalo de confianza*. Aplicando [9.4] se obtiene

$$IC_{\pi_1} = P_1 \pm |Z_{\alpha/2}| \sqrt{P_1(1-P_1)/n} =$$

$$= 0{,}45 \pm 1{,}645 \sqrt{0{,}45(0{,}55)/20} = 0{,}45 \pm 0{,}183 = (0{,}267;\ 0{,}633).$$

Por tanto, estimamos, con una confianza del 95%, que la verdadera proporción de recuperaciones se encuentra entre 0,267 y 0,633.

Este intervalo de confianza, además de dar pistas sobre el verdadero valor del parámetro π_1, sirve para contrastar H_0: como el valor 0,30 propuesto para π_1 en H_0 se encuentra dentro de los limites del intervalo de confianza, se está estimando que el valor 0,30 es uno de los posibles valores del parámetro π_1; por tanto, no es razonable rechazar H_0.

El contraste sobre una proporción con SPSS

Para el *contraste sobre una proporción* (también llamado *prueba binomial*) hemos propuesto dos estrategias. La primera (recomendada para muestras pequeñas) se basa en los estadísticos n_1 y P_1 y utiliza las probabilidades exactas de la distribución binomial; la segunda (recomendada para muestras grandes) se basa en el estadístico Z y utiliza las probabilidades aproximadas de la distribución normal.

El SPSS ofrece ambas soluciones. Con muestras pequeñas ($n \leq 25$), utiliza la distribución binomial para obtener las probabilidades exactas asociadas a los estadísticos n_1 y P_1. Con muestras grandes ($n > 25$) utiliza la distribución normal para obtener las probabilidades asociadas al estadístico Z (es decir, las probabilidades aproximadas asociadas a los estadísticos n_1 y P_1).

Este contraste se encuentra en la opción **Pruebas no paramétricas > Binomial** del menú **Analizar**. La lista de variables del archivo de datos ofrece un listado de todas las variables con formato numérico (no están disponibles las variables con formato de cadena que pueda contener

[4] Como el tamaño muestral no es muy grande, la probabilidad asociada al estadístico Z en la distribución normal (el nivel crítico p) es más parecida a la probabilidad exacta de la distribución binomial si se utiliza la corrección por continuidad. En el ejemplo, el nivel crítico con los estadísticos n_1 y P_1 vale $p = P(n_1 \geq 9) = 0{,}113$, mientras que el nivel crítico con el estadístico Z vale $p = P(Z \geq 1{,}46) = 0{,}0721$. Utilizando la corrección por continuidad se obtiene

$$Z = \frac{9 - 0{,}5 - 20(0{,}30)}{\sqrt{20(0{,}30)(1-0{,}30)}} = \frac{0{,}45 - (0{,}5/20) - 0{,}30}{\sqrt{0{,}30(1-0{,}30)/20}} = 1{,}22$$

en cuyo caso el nivel crítico asociado al estadístico Z vale $p = P(Z \geq 1{,}22) = 0{,}1112$, resultado casi idéntico al obtenido con la distribución binomial (= 0,113).

el archivo de datos). Para obtener la prueba *binomial*: seleccionar una o más variables y trasladarlas a la lista **Contrastar variables**; si se traslada más de una variable, se obtiene un contraste por cada variable.

Las opciones del recuadro **Definir dicotomía** permiten definir qué valores de la variable seleccionada van a utilizarse como categorías: (1) si la variable seleccionada es dicotómica, la opción **Obtener de los datos** deja que sean los propios valores de la variable los que definan la dicotomía; (2) si la variable seleccionada no es dicotómica es necesario dicotomizarla marcando la opción **Punto de corte** e indicando, en el correspondiente cuadro de texto, el valor concreto que se desea utilizar para efectuar el corte: los valores menores o iguales que el punto de corte pasan a formar el primer grupo (categoría) y los valores mayores el segundo[5] (lógicamente, esta segunda opción tiene sentido cuando el nivel de medida de la variable elegida es al menos ordinal).

El cuadro de texto **Proporción de prueba** permite especificar el valor poblacional que se desea contrastar (es decir, el valor asignado a π_1 en la hipótesis nula). Por defecto, se asume que la variable dicotómica seleccionada (o los grupos resultantes de aplicar un punto de corte) sigue el modelo de distribución de probabilidad binomial con $\pi_1 = 0{,}5$; pero ese valor puede cambiarse introduciendo cualquier otro entre 0,001 y 0,999.

De las dos categorías de la variable dicotómica, la que se toma como categoría de referencia es la que corresponde al valor del primer caso válido del archivo de datos. Teniendo esto en cuenta:

1. Si el valor elegido como *proporción de prueba* es 0,5, el SPSS interpreta que el contraste es *bilateral*. En ese caso, si la proporción de casos de la categoría de referencia es mayor que 0,5, el nivel crítico se obtiene multiplicando por dos la probabilidad de encontrar un número de casos igual o mayor que el observado en la categoría de referencia; si la proporción de casos de la categoría de referencia es menor que 0,5 el nivel crítico se obtiene multiplicando por dos la probabilidad de encontrar un número de casos igual o menor que el observado en la categoría de referencia.

2. Si el valor elegido como *proporción de prueba* es distinto de 0,5, el SPSS interpreta que el contraste es *unilateral*. En ese caso, si la proporción de casos de la categoría de referencia es mayor que el valor de la proporción de prueba, el contraste se considera *unilateral derecho* y el nivel crítico se calcula como la probabilidad de obtener un número de casos igual o mayor que el observado en la categoría de referencia; si la proporción de casos de la categoría de referencia es menor que el valor de la proporción de prueba, el contraste se considera *unilateral izquierdo* y el nivel crítico se calcula como la probabilidad de obtener un número de casos igual o menor que el de la categoría de referencia.

---

[5] Esta opción es especialmente útil cuando lo que interesa es contrastar hipótesis sobre la mediana o sobre algún otro cuantil. Es decir, esta opción permite obtener los contrastes conocidos en la literatura estadística como *prueba de los signos* y *prueba de los cuantiles* (ver San Martín y Pardo, 1989, págs. 91-97). Si se desea contrastar, por ejemplo, la hipótesis nula de que la mediana del *salario inicial* vale 25.000 (*prueba de los signos*), puede utilizarse el valor 25.000 como **Punto de corte** y 0,5 (la proporción de casos acumulados hasta la mediana) como valor del contraste en **Contrastar proporción**. Y si se desea contrastar la hipótesis de que el centil 80 del *salario inicial* vale 40.000 (*prueba de los cuantiles*), puede utilizarse 40.000 como **Punto de corte** y 0,80 (la proporción de casos acumulados hasta el centil 80) como valor del contraste en el cuadro de texto **Contrastar proporción**.

Por tanto, las opciones del recuadro **Definir dicotomía** permiten decidir, entre otras cosas, qué tipo de contraste se desea efectuar: sobre una *proporción* (si la variable es dicotómica) o sobre la *mediana* o cualquier otro *cuantil* (si la variable es al menos ordinal).

Ejemplo. El contraste sobre una proporción con SPSS

Este ejemplo muestra cómo utilizar el contraste sobre una proporción o prueba *binomial* con la variable *minoría* (clasificación de minorías) del archivo *Datos de empleados*. Al procedimiento se accede mediante la opción **Pruebas no paramétricas > Binomial** del menú **Analizar**.

Si se asume que el 70% de los habitantes de Estados Unidos es de raza blanca, puede resultar interesante averiguar si ese porcentaje se mantiene en la entidad bancaria a la que se refiere el archivo de datos. Para ello:

- En el cuadro de diálogo principal, seleccionar la variable *minoría* (clasificación de minorías) y trasladarla a la lista **Contrastar variables**.
- Introducir el valor 0,70 en el cuadro de texto **Proporción de prueba** para especificar el valor de la hipótesis nula. En este momento es muy importante tener en cuenta qué valor toma el primer caso del archivo (1 = «sí»; 0 = «no»), pues ese es el valor que el SPSS va a utilizar para identificar a qué categoría de la variable se refiere el valor utilizado como *proporción de prueba*.
- Puesto que la variable es dicotómica, dejar marcada la opción **Obtener de los datos** del recuadro **Definir dicotomía**.
- Pulsar el botón **Opciones** para acceder al subcuadro de diálogo **Prueba binomial: Opciones** y marcar la opción **Descriptivos**; pulsar el botón **Continuar** para volver al cuadro de diálogo principal.

Aceptando estas elecciones, el *Visor* ofrece los resultados que muestran las Tablas 9.1 y 9.2. La Tabla 9.1 contiene el número de casos incluidos en el análisis ($N = 474$), la media (0,22), la desviación típica (0,414) y los valores mínimo y máximo. Dado que se está analizando una variable dicotómica con códigos 1 y 0, la media de la variable indica la proporción de unos (*minoría* = 1 = «sí»). Y la desviación típica es la raíz cuadrada del producto entre la proporción de unos y la proporción de ceros: $(0{,}78 \times 0{,}22)^{1/2} = 0{,}414$ (ver, en el Capítulo 3, la nota a pie de página número 6: si X es una variable dicotómica, entonces: $\sigma_X^2 = \pi_1 - \pi_1^2 = \pi_1(1 - \pi_1)$.

Tabla 9.1. Estadísticos descriptivos

	N	Media	Desviación típica	Mínimo	Máximo
Clasificación de minorías	474	,22	,414	0	1

La Tabla 9.2 comienza identificando la variable utilizada en el contraste y los dos grupos que definen la dicotomía: *grupo 1* (*minoría* = «no») y *grupo 2* (*minoría* = «sí»). Ya hemos señalado que el SPSS toma como categoría de referencia la categoría correspondiente al primer caso válido del archivo de datos: *minoría* = 0 = «no». La *proporción observada* de casos en esa categoría es 370/474 = 0,78 y la *proporción de prueba* es 0,70. Puesto que la proporción de prueba es distinta de 0,5 y la proporción observada de la categoría de referencia es mayor que el valor de prueba (0,78 > 0,70), el SPSS interpreta que el contraste es unilateral derecho y ofrece, como nivel crítico, la probabilidad de obtener, con $n = 474$ y $\pi_1 = 0{,}70$, un número de casos igual o mayor que 370 (el número de casos de la categoría de referencia). Puesto que esa probabilidad (*significación asintótica unilateral*) es menor que 0,0005, se puede rechazar

la hipótesis nula y concluir que la proporción poblacional (*minoría* = «no») es mayor que 0,70. Como el tamaño muestral es mayor que 25, la solución propuesta se basa, no en la distribución binomial, sino en la aproximación normal, lo cual se indica en una nota a pie de tabla.

Tabla 9.2. Prueba binomial

		Categoría	N	Proporción observada	Prop. de prueba	Sig. asintót. (unilateral)
Clasificación de minorías	Grupo 1	No	370	,78	,70	,000[a]
	Grupo 2	Sí	104	,22		
	Total		474	1,00		

a. Basado en la aproximación Z.

La prueba X^2 de Pearson sobre bondad de ajuste

En la investigación aplicada es frecuente encontrar que los investigadores elaboran ideas previas acerca de cómo creen que se comporta la realidad. Estas ideas previas son comunes en todas las disciplinas científicas; se forman a partir de los conocimientos que se van acumulando o, simplemente, observando la realidad y, por lo general, con el objetivo de intentar rellenar alguna laguna de conocimiento.

Gran parte del análisis de datos, particularmente las herramientas más avanzadas, consiste en intentar *ajustar* modelos *teóricos* (que reproducen las ideas de los científicos) a los datos *empíricos* (que reflejan el comportamiento de la realidad) para determinar el grado de parecido o compatibilidad existente entre unos y otros. En este contexto, el concepto de *bondad de ajuste* es un concepto amplio que se refiere al grado de parecido existente entre los pronósticos de un modelo estadístico, cualquiera que éste sea, y los datos que se intenta pronosticar.

Trabajando con una sola variable, los contrastes sobre bondad de ajuste contribuyen, a su manera, a comprobar el grado de parecido existente entre la realidad y esas ideas previas que los científicos se forman acerca de ella. En concreto, permiten comprobar si la *forma* de la distribución de probabilidad *empírica* (la distribución de una variable concreta) se *ajusta* o se parece a una determinada distribución de probabilidad *teórica*.

Estos contrastes pueden utilizarse tanto con variables categóricas como cuantitativas. Con variables dicotómicas se suelen utilizar para valorar el ajuste a una distribución binomial (lo cual equivale al contraste sobre una proporción estudiado en el apartado anterior); con variables politómicas se suelen utilizar para valorar el ajuste a una distribución multinomial; y con variables cuantitativas se suelen utilizar para valorar el ajuste a una distribución normal.

Este apartado se centra en el estudio de cómo una variable politómica (más de dos categorías) se ajusta a una distribución multinomial. Tratamos de dar respuesta a preguntas del tipo: ¿es razonable asumir que, en la población de votantes españoles, el 40% tiene ideología de izquierda, el 20% ideología de centro y el 40% ideología de derecha?, ¿ha cambiado en el último lustro la proporción de fumadores, exfumadores y no fumadores en la población de estudiantes universitarios?, etc. La respuesta a estas preguntas puede obtenerse con una sencilla generalización del estadístico Z utilizado en el apartado anterior.

Supongamos que, de una población cualquiera, se extrae una muestra aleatoria de tamaño n y que cada observación puede ser clasificada en una (y sólo una) de las I categorías de una

variable categórica X. Llamemos i, de forma genérica, a una cualquiera de esas I categorías ($i = 1, 2, ..., I$) y π_i a la probabilidad de que una observación cualquiera sea clasificada en la categoría i ($\pi_i = \pi_1, \pi_2, ..., \pi_I$). Tras clasificar las n observaciones de la muestra tendremos n_1 observaciones en la categoría 1, n_2 observaciones en la categoría 2, ..., n_I observaciones en la categoría I. Estos resultados pueden organizarse en una tabla de frecuencias tal como se muestra en la Tabla 9.3. La tabla también incluye las proporciones observadas ($P_i = n_i / n$) y las teóricas (π_i).

Tabla 9.3. Notación utilizada en tablas de frecuencias

X	n_i	P_i	π_i
1	n_1	$P_1 = n_1 / n$	π_1
2	n_2	$P_2 = n_2 / n$	π_2
...	...	...	...
i	n_i	$P_i = n_i / n$	π_i
...	...	...	...
I	n_I	$P_I = n_I / n$	π_I
	n	1	1

Pearson (1900, 1911) ha ideado una estrategia que permite evaluar si un resultado muestral de estas características se ajusta o no a un determinado tipo de distribución teórica. La estrategia se basa en comparar las **frecuencias observadas** o empíricas (n_i), es decir, de las frecuencias de hecho obtenidas, con las **frecuencias esperadas** o teóricas (m_i), es decir, con las frecuencias que cabría esperar encontrar si la muestra realmente hubiera sido seleccionada de la distribución teórica propuesta[6]:

$$X^2 = \sum_{i=1}^{I} \frac{(n_i - m_i)^2}{m_i} \qquad [9.4]$$

Las frecuencias esperadas se obtienen a partir de las probabilidades teóricas asociadas a cada casilla (ver, en el Capítulo 6, el apartado *Distribución muestral del estadístico proporción*):

$$m_i = n\pi_i \qquad [9.5]$$

El numerador del estadístico de Pearson recoge la diferencia entre cada frecuencia observada y su correspondiente esperada ($n_i - m_i$). A estas diferencias se les llama *residuos*:

$$R_i = n_i - m_i \qquad [9.6]$$

Puesto que el numerador de [9.4] recoge estas diferencias, es claro que el estadístico de Pearson está comparando lo observado con lo esperado. Ahora bien, ¿cómo interpretar una dife-

[6] Este estadístico que aquí estamos llamando X^2 suele encontrarse en la literatura estadística con el nombre *ji*-cuadrado (χ^2). Enseguida veremos que el estadístico X^2 se distribuye según el modelo teórico de probabilidad χ^2 (ver Apéndice 5). Para no confundir el estadístico con su distribución, reservaremos el símbolo X^2 para el estadístico de Pearson y el símbolo χ^2 para la distribución de probabilidad. Por otro lado, recordemos que lo que nosotros estamos llamando *ji*-cuadrado (χ^2) también puede encontrarse en la literatura estadística en español con el nombre *chi*-cuadrado (anglicismo innecesario e incorrecto que, no obstante, nos veremos obligados a utilizar por ser el que aparece en el SPSS).

rencia de, por ejemplo, 5 casos?, ¿es grande o pequeña?, es decir, ¿indica mucha o poca discrepancia entre lo observado y lo esperado? Depende: en una casilla en la que se espera una frecuencia de 10 casos, una diferencia de 5 casos puede ser importante; en una casilla en la que se espera una frecuencia de 1.000 casos, una diferencia de 5 casos parece más bien insignificante. Por tanto, para interpretar correctamente estas diferencias hay que relativizarlas. El estadístico de Pearson las relativiza dividiéndolas entre la frecuencia esperada de la correspondiente casilla. Sumando todas las diferencias de la tabla tras relativizarlas (desde la primera hasta la I-ésima) se obtiene un indicador del grado de discrepancia total existente entre las frecuencias esperadas y las observadas. El problema es que esa suma vale cero; y ésta es la razón por la cual las diferencias del numerador están elevadas al cuadrado.

Lo realmente interesante del estadístico propuesto en [9.4] no es sólo que permite comparar las frecuencias observadas con las esperadas (pues esto podría hacerse de otras maneras), sino que posee *distribución muestral conocida*. Ya sabemos que los estadísticos que utilizamos en los contrastes de hipótesis deben cumplir la doble condición de informar sobre la afirmación establecida en la hipótesis nula y poseer distribución muestral conocida. El estadístico de Pearson se distribuye según el modelo de probabilidad χ^2 (*ji*-cuadrado; ver Apéndice 5) con $I-1$ grados de libertad[7], lo cual representamos mediante

$$X^2 \sim \chi^2_{I-1} \qquad [9.7]$$

Por tanto, la distribución χ^2_{I-1} contiene información sobre las probabilidades asociadas a los diferentes valores de X^2. Y eso es todo lo que hace falta para poder diseñar un contraste basado en la comparación de las frecuencias observadas y las esperadas. El Cuadro 9.2 ofrece un resumen de este contraste.

La Tabla D del Apéndice final ofrece algunos cuantiles de la distribución χ^2. Las cabeceras de las filas indican los grados de libertad de la distribución (gl); las cabeceras de las columnas indican la probabilidad acumulada (proporción de área) que queda por debajo de cada valor χ^2 del interior de la tabla.

Así, por ejemplo, el valor 9,49 correspondiente al cruce entre la cuarta fila ($gl = 4$) y la séptima columna ($p = 0{,}95$) indica que, en la distribución χ^2 con 4 grados de libertad, el valor 9,49 acumula (o sea, deja por debajo o a la izquierda) una proporción de área de tamaño 0,95; es decir, $P(\chi^2_4 < 9{,}49) = F(9{,}49) = 0{,}95$. Para representar este resultado utilizamos la expresión:

$$\chi^2_{4;\,0{,}95} = 9{,}49$$

Los dos subíndices de χ^2 van separados por un punto y coma para evitar confusiones con la coma decimal; el primero de ellos se refiere a los grados de libertad (gl); el segundo, a la proporción de área que deja a la izquierda o que acumula el valor χ^2 (probabilidad acumulada). La tabla únicamente ofrece algunas distribuciones (hasta 30 gl) y algunos cuantiles de esas distribuciones. Para conocer otros valores puede utilizarse un programa informático como el SPSS. Puede consultarse el Apéndice 5 para conocer más detalles sobre la distribución *ji*-cuadrado y sobre la forma de utilizar tanto las tablas como el SPSS para conocer sus probabilidades.

[7] La aproximación de X^2 a χ^2 es tanto mejor cuanto mayor es el tamaño muestral. No existe una regla que indique cómo de grande debe ser el tamaño muestral para que la aproximación sea lo bastante buena, pero sí se sabe que la aproximación es tanto mejor cuanto mayores son las frecuencias esperadas (ver Apéndice 9).

Cuadro 9.2. Resumen del contraste sobre bondad de ajuste con la prueba X^2 de Pearson

1. *Hipótesis*:

   $H_0: f(n_i) = M(n; \pi_1, \pi_2, ..., \pi_I)$.
   Es decir, la función de probabilidad de la variable n_i es multinomial con probabilidades teóricas $\pi_1, \pi_2, ..., \pi_I$ asociadas a cada categoría i de la variable X (la variable n_i se refiere a los posibles resultados que pueden originarse al clasificar n casos en las i categorías de la variable X).

   $H_1: f(n_i) \neq M(n; \pi_1, \pi_2, ..., \pi_I)$.
   Es decir, la función de probabilidad de la variable n_i no es la propuesta en H_0.

2. *Supuestos*: tenemos una muestra aleatoria de n observaciones (n ensayos) clasificada en las I categorías exclusivas y exhaustivas de una variable categórica. La probabilidad de que una observación pertenezca a cada una de las categorías de la variable se mantiene constante en los n ensayos (es decir, las n observaciones son independientes entre sí).Y no más del 20% de las frecuencias esperadas[8] (m_i) son menores que 5.

3. *Estadístico del contraste* (ver ecuación [9.4]):

   $$X^2 = \sum_{i=1}^{I} \frac{(n_i - m_i)^2}{m_i}$$

4. *Distribución muestral*: X^2 se aproxima a χ^2 con $I - 1$ grados de libertad. La aproximación es tanto mejor cuanto mayores son las frecuencias esperadas de las casillas (ver Apéndice 9).

5. *Zona crítica*: $X^2 \geq \chi^2_{I-1;\, 1-\alpha}$.

6. *Regla de decisión*: se rechaza H_0 si el estadístico X^2 cae en la zona crítica; en caso contrario, se mantiene.

7. *Nivel crítico* (valor p): $p = P(X^2 \geq X^2_h)$, siendo X^2_h el valor concreto que toma X^2.

8. *Intervalo de confianza* para la proporción teórica de cada casilla.
   Siendo $P_i = n_i/n$ (ver Tabla 9.3) y teniendo en cuenta que la frecuencia de cada casilla se distribuye según el modelo de probabilidad binomial:

   $$IC_{\pi_i} = P_i \pm |Z_{\alpha/2}| \sqrt{P_i(1 - P_i)/n} \qquad [9.8]$$

   Si se rechaza H_0, este intervalo de confianza permite determinar en qué categorías de la variable falla el ajuste. En concreto, decidiremos que el ajuste se rompe en una categoría cualquiera i cuando el intervalo construido para esa categoría a partir de P_i no incluya el valor de la correspondiente proporción teórica π_i.

[8] Esta afirmación sobre las frecuencias esperadas, más que una condición necesaria para poder aplicar el estadístico X^2, es una recomendación relacionada con la conveniencia de no utilizar el estadístico X^2 con muestras muy pequeñas. En el Apéndice 9, en el apartado *Supuestos de la prueba X^2 de Pearson.*, se ofrece una explicación algo más detallada sobre este supuesto (o, mejor, sobre la conveniencia de no utilizar el estadístico X^2 de Pearson cuando las frecuencias esperadas son demasiado pequeñas).

Ejemplo. La prueba X^2 de Pearson sobre bondad de ajuste

Por información recogida en varios estudios se sabe que, en la población de madrileños mayores de 25 años, la proporción de fumadores, exfumadores y no fumadores es de 0,30, 0,12 y 0,58, respectivamente. Se desea averiguar si, en la población de jóvenes con edades comprendidas entre los 15 y los 25 años, se reproduce esa misma pauta. Para ello, se ha seleccionado una muestra aleatoria de 250 jóvenes en la que se han encontrado 88 fumadores, 12 exfumadores y 150 no fumadores. ¿Qué puede concluirse, con un nivel de confianza de 0,95?

Tenemos una variable politómica (categórica con tres categorías) y debemos averiguar si las frecuencias observadas (las frecuencias de hecho obtenidas en la muestra) se parecen o no a las esperadas (a las que cabría esperar si la muestra de universitarios perteneciera a una población con las proporciones teóricas propuestas).

1. *Hipótesis*:

   $H_0 : f(n_i) = M(\pi_{\text{fumadores}} = 0{,}30;\ \pi_{\text{exfumadores}} = 0{,}12;\ \pi_{\text{no fumadores}} = 0{,}58)$.
   $H_0 : f(n_i) \neq M(\pi_{\text{fumadores}} = 0{,}30;\ \pi_{\text{exfumadores}} = 0{,}12;\ \pi_{\text{no fumadores}} = 0{,}58)$.

2. *Supuestos*: 250 sujetos aleatoriamente seleccionados se han clasificado en las $I = 3$ categorías exclusivas y exhaustivas de una variable politómica. La probabilidad teórica asociada a cada categoría (0,30, 0,12 y 0,58) se mantiene constante durante todo el proceso de clasificación. No existen frecuencias esperadas menores que 5.

3. *Estadístico del contraste*: la Tabla 9.4 muestra las frecuencias observadas (n_i), las proporciones teóricas (π_i) y las esperadas (m_i). Las frecuencias esperadas se han obtenido aplicando la ecuación [9.5]; por ejemplo, la frecuencia esperada correspondiente a la categoría *fumadores* se ha obtenido mediante: $m_{\text{fumadores}} = n\,\pi_{\text{fumadores}} = 250\,(0{,}30) = 75$.

Tabla 9.4. Frecuencias observadas y esperadas

X = *Tabaquismo*	n_i	π_i	m_i
1 = Fumadores	88	0,30	75
2 = Exfumadores	12	0,12	30
3 = No fumadores	150	0,58	145
	250	1	250

$$X^2 = \frac{(88-75)^2}{75} + \frac{(12-30)^2}{30} + \frac{(150-145)^2}{145} = 13{,}23$$

4. *Distribución muestral*: X^2 se aproxima a χ^2 con $I - 1 = 3 - 1 = 2$ grados de libertad: χ^2_2.

5. *Zona crítica*: $X^2 \geq \chi^2_{2;\,0{,}95} = 5{,}99$.

6. *Decisión*: como $X^2 = 13{,}23$ es mayor que el punto crítico 5,99, se rechaza H_0. Puede concluirse, por tanto, que las frecuencias observadas no se ajustan a las que se derivan de las proporciones propuestas en la hipótesis nula. Es decir, no parece probable que la muestra seleccionada proceda de una población donde las proporciones de fumadores, exfumadores y no fumadores sean, respectivamente 0,30, 0,12 y 0,58.

7. *Nivel crítico* (valor p): $p = P(X^2 \geq 13{,}23) < 0{,}005$.

8. *Intervalo de confianza*: la Tabla 9.5 muestra los límites inferior y superior del intervalo de confianza construido para cada categoría de la variable *tabaquismo*. Estos intervalos se han calculado aplicando la ecuación [9.8]. Así, por ejemplo, el intervalo correspondiente a la categoría *fumadores* se ha obtenido de la siguiente manera:

$$IC_{\pi_{\text{fumadores}}} = 0{,}352 \pm 1{,}96\sqrt{0{,}352(1-0{,}352)/250} = (0{,}293;\ 0{,}411)$$

Tabla 9.5. Frecuencias observadas y esperadas

X = *Tabaquismo*	n_i	$P_i = n_i / n$	L_i	L_s	π_i
1 = Fumadores	88	88/250 = 0,352	0,293	0,411	0,30 incluida
2 = Exfumadores	12	123/250 = 0,048	0,022	0,074	0,12 no incluida
3 = No fumadores	150	150/250 = 0,600	0,540	0,660	0,58 incluida

Los resultados de la tabla indican que únicamente la proporción teórica de la categoría *exfumadores* (0,12) queda fuera de los límites de confianza obtenidos. Por tanto, puede concluirse que únicamente en esa categoría se rompe el ajuste entre lo observado y lo esperado. Es decir, no puede afirmarse que la proporción de fumadores y no fumadores en la población de jóvenes sea distinta de esas mismas proporciones en la población general; pero sí existe evidencia de que la proporción de exfumadores es significativamente menor en la población de jóvenes que en la población general.

La prueba X^2 de Pearson sobre bondad de ajuste con SPSS

El estadístico de Pearson sobre bondad de ajuste se encuentra en la opción **Pruebas no paramétricas > Chi-cuadrado** del menú **Analizar**. El procedimiento sólo permite utilizar variables con formato numérico (no están disponibles las variables con formato de cadena). Para contrastar la hipótesis de bondad de ajuste referida a una variable categórica: trasladar la variable a la lista **Contrastar variables** (si se selecciona más de una variable, el SPSS ofrece tantos contrastes como variables se seleccionen).

El recuadro **Rango esperado** permite decidir qué rango de valores de la variable seleccionada deben ser incluidos en el análisis. Si se elige la opción **Obtener de los datos**, cada valor distinto de la variable se considera una categoría para el análisis. Si se elige la opción **Usar rango especificado**, sólo se tienen en cuenta los valores comprendidos entre los límites que se especifiquen en los cuadros de texto **Inferior** y **Superior** (los valores no incluidos entre estos límites se excluyen del análisis).

Las opciones del recuadro **Valores esperados** sirven para indicar cuál debe ser el valor de las frecuencias esperadas. Si se elige la opción **Todas las categorías iguales**, las frecuencias esperadas se calculan dividiendo el número total de casos válidos entre el número de categorías de la variable, lo cual implica contrastar la hipótesis nula de que la variable seleccionada se ajusta a una distribución multinomial donde todas las categorías tienen asociada la misma

probabilidad teórica. La opción **Valores** permite definir frecuencias esperadas concretas. Al utilizar esta opción es importante tener en cuenta que los valores que se introducen se interpretan como proporciones, no como frecuencias absolutas. Deben introducirse tantos valores como categorías: el SPSS divide cada valor por la suma de todos los valores. Así, por ejemplo, si una variable tiene tres categorías y se introducen los números enteros 6, 1 y 3, el SPSS interpreta que la frecuencia esperada de la primera categoría es 6/10 del número de casos válidos, la de la segunda categoría 1/10 del número de casos válidos, y la de la tercera categoría 3/10 del número de casos válidos. De esta forma, resulta fácil definir, por ejemplo, las frecuencias esperadas correspondientes a una distribución binomial o multinomial. El orden en el que se introducen los valores es muy importante: el procedimiento hace corresponder la secuencia introducida con las categorías de la variable cuando éstas se encuentran ordenadas de forma ascendente (es decir, de menor a mayor).

El botón **Opciones** ofrece la posibilidad de elegir algunos estadísticos descriptivos y decidir qué tratamiento se desea dar a los valores perdidos. La opción **Descriptivos** ofrece el número de casos válidos, la media, la desviación típica, y los valores mínimo y máximo; la opción **Cuartiles** ofrece los percentiles 25, 50 y 75 (lógicamente, estos estadísticos sólo tienen interés si la variable es cuantitativa; y puesto que el estadístico de Pearson se utiliza con variables categóricas –pues con variables cuantitativas se suele utilizar la prueba de Kolmogorov-Smirnov–, normalmente no será necesario solicitarlos).

Por lo que se refiere a los **Valores perdidos**, si se elige la opción **Excluir casos según prueba** se excluyen de cada contraste los casos con valor perdido en la variable que se está contrastando (es la opción por defecto); si se elige la opción **Excluir casos según pareja** se excluyen de todos los contrastes solicitados los casos con algún valor perdido en cualquiera de las variables seleccionadas.

Ejemplo. La prueba X^2 de Pearson sobre bondad de ajuste con SPSS

Este ejemplo muestra cómo realizar con el SPSS el mismo contraste que en el ejemplo anterior. Recordemos que se trata de averiguar si, en la población de jóvenes con edades comprendidas entre los 15 y los 25 años, la variable *tabaquismo* se distribuye de igual forma que en la población general.

Lo que ocurre en la población general es que hay un 30% de fumadores, un 12% de exfumadores y un 58% de no fumadores. Y al seleccionar una muestra aleatoria de jóvenes entre 15 y 25 años se han encontrado 88 fumadores, 12 exfumadores y 150 no fumadores. ¿Qué puede concluirse, con un nivel de confianza de 0,95?

Lo primero que necesitamos hacer es reproducir en el *Editor de datos* del SPSS los datos de la Tabla 9.4. La Figura 9.1 muestra cómo hacerlo (ver Pardo y Ruiz, 2009, págs. 171-173).

Figura 9.1. Datos de la Tabla 9.4 reproducidos en el *Editor de datos*

	tabaco	n_tabaco
1	1	88
2	2	12
3	3	150

Hemos creado dos variables: *tabaco* (con etiqueta *Tabaquismo* y códigos 1, 2 y 3) y *n_tabaco* (con las frecuencias observadas). Si se desea, se pueden asignar etiquetas a los códigos (categorías) de la variable *tabaco*: 1 = «Fumadores», 2 = «Exfumadores», 3 = «No fumadores». Ponderando ahora el archivo con la variable *n_tabaco* (con la opción **Ponderar** del menú **Datos**), los 3 casos del archivo se convierten en los 250 de la Tabla 9.4.

Para contrastar la hipótesis de bondad de ajuste con el estadístico X^2 de Pearson (prueba *ji*-cuadrado o prueba *chi*-cuadrado):

- ➤ Seleccionar la opción **Pruebas no paramétricas** > **Chi-cuadrado** del menú **Analizar** y, en el cuadro de diálogo principal, seleccionar la variable *tabaco* (tabaquismo) y trasladarla a la lista **Contrastar variables**.
- ➤ En el recuadro **Valores esperados**, seleccionar la opción **Valores** e introducir los valores 30, 12 y 58 utilizando el correspondiente cuadro de texto y el botón **Añadir.**

  Puesto que la variable seleccionada es categórica, no tiene sentido solicitar los descriptivos (media, desviación típica, cuartiles, etc.) disponibles en el botón **Opciones**.

Aceptando estas elecciones, el *Visor* ofrece los resultados que muestran las Tablas 9.6 y 9.7. La Tabla 9.6 contiene las frecuencias observadas (*N observado*) y las esperadas (*N esperado*), así como las diferencias entre ambas (*residuo*). La Tabla 9.7 ofrece la información necesaria para tomar una decisión sobre la hipótesis nula de bondad de ajuste: el valor del estadístico de Pearson (*chi-cuadrado* = 13,226), sus grados de libertad (*gl* = 2) y su nivel crítico (*sig. asintótica* = 0,001). Puesto que el nivel crítico es menor que 0,05, se puede rechazar la hipótesis nula de bondad de ajuste y concluir que las frecuencias de las categorías de la variable *tabaquismo* no se ajustan a la distribución propuesta en la hipótesis nula (0,30, 0,12, 0,58).

Tabla 9.6. Frecuencias observadas, esperadas y residuos (variable *tabaco*)

	N observado	N esperado	Residuo
Fumadores	88	75,0	13,0
Exfumadores	12	30,0	-18,0
No fumadores	150	145,0	5,0
Total	250		

Tabla 9.7. Estadístico *chi*-cuadrado

	Tabaquismo
Chi-cuadrado[a]	13,226
gl	2
Sig. asintót.	,001

a. 0 casillas (0%) tienen frecuencias esperadas menores que 5. La frecuencia esperada mínima es 30,0.

En una nota a pie de tabla se indica el número y porcentaje de casillas con frecuencias esperadas menores que 5. Recordemos que, puesto que el estadístico de Pearson se aproxima a la distribución *chi*-cuadrado tanto mejor cuanto mayor es el tamaño muestral, suele asumirse (siguiendo la recomendación de Cochran, 1952) que, si existen frecuencias esperadas menores que 5, éstas no deben superar el 20% del total de frecuencias de la tabla (ver Apéndice 9).

El contraste sobre una media (prueba *T* para una muestra)

El contraste de hipótesis sobre una media sirve para tomar decisiones acerca del valor poblacional de la media de una variable. Permite responder a preguntas del tipo: ¿puede afirmarse que el cociente intelectual medio de un determinado colectivo es mayor que 100?, ¿se parece la media estandarizada que se obtiene con una nueva prueba de rendimiento a la que se viene obteniendo tradicionalmente?, ¿es cierto que el peso medio de los recién nacidos de madres fumadoras no alcanza los 2,5 kg?, etc. Todas estas preguntas coinciden en dirigir la atención al *centro* de *una variable cuantitativa*.

Ya hemos estudiado en el Capítulo 5 la distribución muestral del estadístico media. Y, al describir la lógica general del contraste de hipótesis en el capítulo anterior, hemos utilizado, entre otros, ejemplos referidos a la media. Sabemos que la media $\bar{Y}$ de una muestra aleatoria de tamaño n extraída de una población normal $N(\mu_Y,\ \sigma_Y)$ es un estadístico (una variable) que se distribuye normalmente con parámetros μ_Y y $\sigma_{\bar{Y}}$ (ver en el Capítulo 6, el apartado *Distribución muestral del estadístico media*):

$$\bar{Y} \quad \sim \quad N(\mu_Y,\ \sigma_{\bar{Y}}) \qquad [9.9]$$

con $\sigma_{\bar{Y}} = \sigma_Y / \sqrt{n}$. Sabemos también, por el teorema del límite central, que, aun no siendo normal la población original, el estadístico $\bar{Y}$ tiende a distribuirse normalmente $N(\mu_Y,\ \sigma_{\bar{Y}})$ a medida que el tamaño de la muestra va aumentando. Y también sabemos, por último, que, bajo las mencionadas circunstancias, se verifica (ver ecuación [5.7]):

$$Z \;=\; \frac{\bar{Y} - \mu_Y}{\sigma_{\bar{Y}}} \;=\; \frac{\bar{Y} - \mu_Y}{\sigma_Y / \sqrt{n}} \quad \sim \quad N(0,1) \qquad [9.10]$$

Por tanto, es posible utilizar la transformación Z junto con la distribución normal tipificada para conocer las probabilidades asociadas a los diferentes valores del estadístico $\bar{Y}$. Lo cual significa que tenemos todo lo necesario para contrastar hipótesis sobre el parámetro μ.

Pero la utilidad del estadístico Z propuesto en [9.10] es bastante escasa cuando se llevan a cabo estudios reales. Generalmente, si se conoce la desviación típica de una población, también se conocerá su media y, por tanto, no será necesario hacer ningún tipo de inferencia sobre ella. Y, si conociendo ambos parámetros, se desea averiguar si la media ha cambiado como consecuencia de, por ejemplo, algún tipo de intervención, lo razonable será asumir que también la varianza habrá podido cambiar y que, por tanto, habrá dejado de ser conocida.

Estas consideraciones sugieren que, al contrastar hipótesis sobre una media, lo habitual es que los parámetros poblacionales (tanto μ como σ) sean desconocidos. Y cuando esto es así, las cosas cambian. Recordemos (ver, en el Capítulo 5, el apartado *Distribución muestral del estadístico media*) que, si de una población en la que la variable Y se distribuye normalmente se extrae una muestra aleatoria de tamaño n y se calcula el estadístico $\bar{Y}$, se verifica que

$$T \;=\; \frac{\bar{Y} - \mu_Y}{\hat{\sigma}_{\bar{Y}}} \;=\; \frac{\bar{Y} - \mu_Y}{S_Y / \sqrt{n}} \quad \sim \quad t_{n-1} \qquad [9.11]$$

Es decir, al sustituir σ_Y por S_Y, la transformación resultante ya no se distribuye $N(0,1)$, como ocurre con la transformación Z, sino según el modelo de probabilidad t de Student con $n-1$

grados de libertad[9]. Por tanto, cuando se desconoce el parámetro σ_Y, es posible conocer las probabilidades asociadas al estadístico $\overline{Y}$ mediante la transformación T y la distribución t de Student. Y esto es todo lo que hace falta para diseñar un contraste sobre el parámetro μ_Y cuando se desconoce el parámetro σ_Y. El Cuadro 9.3 ofrece un resumen de este contraste.

Independencia y normalidad

Para poder afirmar que la distribución muestral del estadístico T se aproxima a la distribución teórica t de Student se ha impuesto, como punto de partida, que la muestra sea aleatoria y la población muestreada normal. En ese punto de partida hay implícitas dos condiciones: *independencia* y *normalidad*. Recordemos que, a las condiciones que deben darse para que un estadístico se distribuya como se dice que se distribuye, las llamamos *supuestos*.

El supuesto de *independencia* se refiere a que las observaciones deben ser independientes entre sí, lo cual significa que el resultado de una observación no debe condicionar el resultado de ninguna otra (cuando las observaciones son independientes, conocer el resultado de una de ellas no nos dice nada acerca del resultado de las demás). Ésta es la razón por la cual las observaciones se seleccionan aleatoriamente: la aleatoriedad del muestreo es lo que garantiza la independencia entre las observaciones. El incumplimiento de este supuesto puede alterar seriamente la distribución muestral del estadístico T y esto puede llevar a tomar decisiones equivocadas (ver, por ejemplo, Kenny y Judd, 1986).

El supuesto de *normalidad* se refiere a que la variable analizada se distribuye normalmente en la población muestreada. Este supuesto es muy importante cuando se trabaja con muestras pequeñas, pero va perdiendo importancia conforme va aumentando el tamaño muestral. Para ayudar a decidir cuándo una muestra es lo bastante grande como para no tener que preocuparse por el supuesto de normalidad, Moore (2007, págs. 447-448), basándose en los resultados obtenidos en varios trabajos de simulación (Posten, 1979; Pearson y Please, 1975; ver también Boos y Hughes-Oliver, 2000; Sawilowsky y Blair, 1992), hace la siguiente recomendación: (1) para utilizar la prueba T con tamaños muestrales menores de 15 es indispensable que los datos se distribuyan de forma aproximadamente normal[10] (sin asimetrías evidentes y sin valores atípicos); (2) con tamaños muestrales comprendidos entre 15 y 40 puede utilizarse la prueba T siempre que los datos no se distribuyan de forma muy asimétrica y no existan valores atípicos; (3) con tamaños muestrales por encima de 40, la prueba T puede utilizarse incluso aunque la distribución de los datos sea fuertemente asimétrica y existan valores atípicos. En el apartado *Relación entre las distribuciones t y χ^2* del Apéndice 9 se discuten algunos aspectos relacionados con el supuesto de normalidad.

---

[9] En el Apéndice 5 se ofrece una descripción de la distribución t; y en el Apéndice 9 se explica con más detalle la diferencia existente entre las transformaciones Z y T. Si el tamaño muestral es lo bastante grande, el estadístico Z propuesto en [9.10] y el estadístico T propuesto en [9.11] ofrecen resultados muy parecidos. Esto significa que, a medida que el tamaño muestral va aumentando, va resultando irrelevante el hecho de que el parámetro σ_Y sea conocido o desconocido: al aumentar el tamaño muestral, los posibles valores de S_Y se van aproximando más y más al valor de σ_Y, y la distribución de T se va pareciendo más y más a la distribución de Z (ver Apéndice 9). Por ejemplo, el percentil 95 de la distribución normal estandarizada vale 1,645; y en la distribución t con $gl=10$ vale 1,812; con $gl=50$, 1,676; con $gl=100$, 1,660; etc. Por tanto, si el tamaño muestral es lo bastante grande, siempre resulta posible utilizar la distribución normal para conocer las probabilidades asociadas a la media, tanto si se conoce σ_Y como si no.

[10] En el caso de que sea necesario trabajar con muestras pequeñas y distribuciones no normales, existen procedimientos alternativos a la prueba T (prueba de Wilcoxon, prueba de los signos) que se estudian en el segundo volumen.

Cuadro 9.3. Resumen del contraste sobre una media (prueba *T* para una muestra)

1. *Hipótesis*:
   a. Contraste bilateral: H_0: $\mu_Y = k_0$; H_1: $\mu_Y \neq k_0$.
   b. Contraste unilateral derecho: H_0: $\mu_Y \leq k_0$; H_1: $\mu_Y > k_0$.
   c. Contraste unilateral izquierdo: H_0: $\mu_Y \geq k_0$; H_1: $\mu_Y < k_0$.
   (k_0 se refiere al valor concreto de μ_Y que interesa contrastar).
2. *Supuestos*: muestra aleatoria de tamaño *n* extraída de una población normal (el supuesto de *normalidad* va perdiendo importancia –ver apartado anterior– conforme el tamaño muestral va aumentando).
3. *Estadístico del contraste* (ver ecuación [9.11]):

$$T = \frac{\bar{Y} - \mu_Y}{S_Y / \sqrt{n}}$$

4. *Distribución muestral*: *T* se distribuye según *t* con $n-1$ grados de libertad (t_{n-1}).
5. *Zona crítica*:
   a. Contraste bilateral: $T \leq t_{n-1;\,\alpha/2}$ y $T \geq t_{n-1;\,1-\alpha/2}$.
   b. Contraste unilateral derecho: $T \geq t_{n-1;\,1-\alpha}$.
   c. Contraste unilateral izquierdo: $T \leq t_{n-1;\,\alpha}$.
6. *Regla de decisión*: se rechaza H_0 si el estadístico del contraste cae en la zona crítica; en caso contrario, se mantiene.
7. *Nivel crítico* (valor *p*):
   a. Contraste bilateral: $p = 2\,[P(T \geq |T_h|)]$, siendo T_h el valor muestral concreto que toma el estadístico *T*.
   b. Contraste unilateral derecho: $p = P(T \geq T_h)$.
   c. Contraste unilateral izquierdo: $p = P(T \leq T_h)$.
8. *Intervalo de confianza* (ver ecuación [7.11]): $IC_{\mu_Y} = \bar{Y} \pm t_{n-1;\,1-\alpha/2}\, S_Y / \sqrt{n}$.

Ejemplo. El contraste sobre una media (prueba T para una muestra)

Supongamos que en un centro de educación especial se utiliza, para estimular la comprensión lectora de los niños, un método con el que se viene obteniendo una media de 6 en una determinada prueba de comprensión lectora. Un educador especialista en problemas de lectura ofrece al centro la posibilidad de utilizar un nuevo método que, según él, es más económico y eficiente. El centro estaría dispuesto a adoptar el nuevo método siempre que el rendimiento en comprensión lectora no fuera inferior al que se viene obteniendo con el método actual. Para valorar esta circunstancia, se aplica el nuevo método a una muestra aleatoria de 20 niños.

Tras la instrucción, se pasa la prueba de comprensión lectora y se obtiene una media de 5 y una desviación típica de 1,3. Con este resultado, y considerando que en la distribución de los datos no se da ni fuerte asimetría ni valores atípicos, ¿qué decisión debe tomarse? ($\alpha = 0{,}05$).

1. *Hipótesis*:

   H_0: $\mu_Y \geq 6$

   H_1: $\mu_Y < 6$ (contraste unilateral izquierdo).

2. *Supuestos*: tenemos una muestra aleatoria de 20 puntuaciones en comprensión lectora procedentes de una población que no sabemos si es o no normal; no obstante, puesto que se nos dice que en la distribución de los datos no existe ni fuerte asimetría ni valores atípicos y el tamaño muestral es mayor de 15, podemos utilizar la prueba *T* sin necesidad de asumir normalidad.

3. *Estadístico del contraste*:

$$Z = \frac{\bar{Y} - \mu_Y}{S_Y/\sqrt{n}} = \frac{5-6}{1{,}3/\sqrt{20}} = \frac{-1}{0{,}29} = -3{,}45$$

4. *Distribución muestral*: *T* se distribuye según $t_{n-1} = t_{19}$.

5. *Zona crítica* (contraste unilateral izquierdo): $T \leq t_{19;\,0{,}05} = -1{,}729$.

6. *Decisión*: como $-3{,}45 < -1{,}729$, se rechaza H_0. Por tanto, puede concluirse que el promedio obtenido con el nuevo método es significativamente menor que el que se viene obteniendo con el método actual; en consecuencia, no parece haber justificación para adoptar el nuevo método.

7. *Nivel crítico* (valor *p*)[11]: $p = P(T \leq -3{,}45) < 0{,}005$.

8. *Intervalo de confianza* (con $S_Y/\sqrt{n} = 0{,}29$; denominador del estadístico *Z*):

$$IC_{\mu_Y} = \bar{Y} \pm t_{19;\,0{,}975}\, S_Y/\sqrt{n} = 5 \pm 2{,}093\,(0{,}29) = 5 \pm 0{,}61 = (4{,}39;\ 5{,}61)$$

   Podemos estimar, con una confianza del 95%, que el promedio poblacional que se obtiene en la prueba de comprensión lectora con el nuevo método de enseñanza se encuentra entre 4,39 y 5,61.

El contraste sobre una media (prueba *T* para una muestra) con SPSS

El contraste sobre una media se encuentra en la opción **Comparar medias > Prueba *T* para una muestra** del menú **Analizar**. La lista de variables del cuadro de diálogo principal sólo contiene las variables *numéricas* del archivo de datos (no es posible, por tanto, utilizar variables con formato de *cadena*).

[11] La tabla de la distribución *t* que aparece en el Apéndice final no es lo bastante completa como para poder obtener a partir de ella el nivel crítico exacto. Sin embargo, esto no debe ser considerado un inconveniente serio pues, por lo general, para tomar una decisión sobre la hipótesis nula es suficiente con saber si el nivel crítico (*p*) es mayor o menor que el nivel de significación establecido (α).

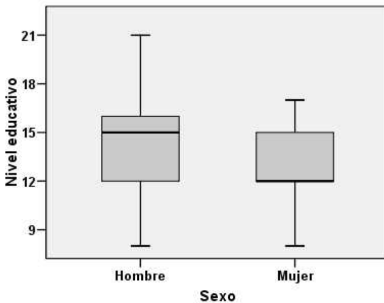
Para llevar a cabo un contraste con las especificaciones que el procedimiento tiene establecidas por defecto: (1) seleccionar la variable cuya media poblacional se desea contrastar y trasladarla a la lista **Variables para contrastar** (puede seleccionarse más de una variable; cada variable seleccionada genera un contraste); (2) introducir en el cuadro de texto **Valor de prueba** el valor poblacional concreto que se desea contrastar, es decir, el valor asignado a μ en la hipótesis nula (este valor se aplica a todas las variables seleccionadas).

Además de algunos descriptivos, el procedimiento ofrece un resumen del contraste y un intervalo de confianza. Las opciones del procedimiento permiten controlar algunos aspectos del análisis. La opción **Intervalo de confianza: *k*%** permite establecer, en escala porcentual, el *nivel de confianza* $(1 - \alpha)$ con el que se desea construir el intervalo de confianza para la diferencia entre el valor de la media muestral y el valor propuesto para la media poblacional. El valor de *k* es, por defecto, 95, pero es posible elegir cualquier otro valor comprendido entre 0,01 y 99,99. Las opciones del recuadro **Valores perdidos** permite elegir el tratamiento que se desea dar a los valores perdidos: la opción **Excluir casos según análisis** excluye de cada análisis (de cada contraste) los casos con valor perdido en la variable concreta que se está contrastando; la opción **Excluir casos según lista** excluye de todos los análisis los casos con algún valor perdido en una cualquiera de las variables seleccionadas.

Ejemplo. El contraste sobre una media (prueba T para una muestra) con SPSS

Este ejemplo muestra cómo contrastar hipótesis sobre una media con SPSS. Sabemos que la media de la variable *educ* (nivel educativo) del archivo *Datos de empleados* vale aproximadamente 13,5 años[12] (este archivo es el mismo que venimos utilizando en otros ejemplos y se encuentra en la misma carpeta en la que está instalado el SPSS). Imaginemos que deseamos averiguar si ese promedio es asumible tanto en la población de hombres como en la de mujeres. Un análisis descriptivo preliminar mediante un diagrama de cajas (ver Figura 9.2) indica que el nivel educativo medio de los hombres supera el promedio global (13,5) mientras que el nivel educativo medio de las mujeres no lo alcanza: la media de los hombres se sitúa cerca de 15 y la de las mujeres cerca de 12.

Figura 9.2. Diagrama de cajas del *nivel educativo* en *hombres* y en *mujeres*



[12] Debe tenerse en cuenta que, en este archivo, el nivel educativo no es una variable categórica, sino una variable cuantitativa medida como *años de formación académica*.

La pregunta a la que tratamos de dar respuesta es la siguiente: ¿es razonable asumir, a partir de la información muestral disponible, que tanto en la población de hombres como en la de mujeres el nivel educativo medio es de 13,5 años? Es decir, ¿es razonable asumir que las diferencias observadas entre las medias muestrales y la media poblacional no va más allá de la esperable por las fluctuaciones propias del azar muestra? Para dar respuesta a esta pregunta debemos llevar a cabo dos contrastes[13], uno con cada población, planteando en ambos casos la hipótesis nula de que la media poblacional del nivel educativo vale 13,5 años. Para ello:

- Seleccionar la opción **Segmentar archivo** del menú **Datos**, marcar la opción **Comparar los grupos**, trasladar la variable *sexo* al cuadro **Grupos basados en** y pulsar el botón **Aceptar** (esta acción sirve para dividir el archivo en dos grupos –hombres y mujeres– y, a partir de ahí, poder trabajar con cada grupo por separado).
- Seleccionar la opción **Comparar medias** > **Prueba *T* para una muestra** del menú **Analizar** y trasladar la variable *educ* (nivel educativo) a la lista **Variables para contrastar**.
- Introducir el valor 13,5 en el cuadro de texto **Valor de prueba**.

Aceptando estos valores, el *Visor de resultados* ofrece la información que muestran las Tablas 9.8 y 9.9. Puesto que hemos segmentado el archivo utilizando la variable *sexo*, las tablas ofrecen resultados separados para el grupo de hombres y para el de mujeres.

La Tabla 9.8 incluye, para cada grupo, el número de casos válidos sobre el que se basa cada contraste (258 hombres y 216 mujeres), la media observada en cada grupo (14,43 en hombres y 12,37 en mujeres), la desviación típica insesgada (2,979 en hombres y 2,319 en mujeres) y el error típico de la media (0,185 en hombres y 0,158 en mujeres; recordemos que el error típico de la media es el denominador del estadístico T y que se obtiene dividiendo la desviación típica insesgada entre la raíz cuadrada del número de casos).

Tabla 9.8. Estadísticos descriptivos del procedimiento *Prueba T para una muestra*

Nivel educativo

Sexo	N	Media	Desviación típica.	Error típico de la media
Hombre	258	14,43	2,979	,185
Mujer	216	12,37	2,319	,158

La Tabla 9.9 ofrece un resumen de los dos contrastes solicitados. El encabezamiento de la tabla recuerda cuál es el valor propuesto para la media poblacional en la hipótesis nula (*valor de prueba* = 13,5). No olvidar que el valor de prueba es el mismo para los dos contrastes solicitados. La hipótesis nula que se está poniendo a prueba en el primer contraste es $\mu_{\text{hombres}} = 13{,}5$; y en el segundo, $\mu_{\text{mujeres}} = 13{,}5$. Las tres primeras columnas de la tabla contienen el valor de los estadísticos[14] T ($t_h = 5{,}015$; $t_m = -7{,}159$), sus grados de libertad ($gl_h = 257$; $gl_m = 215$) y

[13] Es importante reparar en el hecho de que no estamos interesados en comparar los promedios de ambas poblaciones (esto es algo que aprenderemos a hacer en el próximo capítulo). De momento, en este ejemplo, sólo se está intentando averiguar si es o no razonable pensar que el nivel educativo medio de ambas poblaciones vale 13,5 años.

[14] En el SPSS, como en otros muchos sitios, el estadístico al que nosotros estamos llamando T se suele representar con letras minúsculas. Nosotros, para diferenciar entre el estadístico y su distribución, reservamos la letra mayúscula T para el estadístico (recordemos que las variables las representamos siempre con letras mayúsculas) y la letra minúscula t para la distribución a la que se aproxima el estadístico.

el nivel crítico bilateral o valor p (*sig. bilateral* < 0,0005 en ambos casos[15]). El nivel crítico unilateral puede obtenerse dividiendo entre 2 el bilateral. Recordemos que el nivel crítico indica el grado de compatibilidad existente entre el valor poblacional propuesto para la media y la información muestral disponible. Y la regla de decisión adoptada dice que si el nivel crítico es pequeño (generalmente menor que 0,05), puede afirmarse que los datos son incompatibles con la hipótesis nula de que el verdadero valor de la media poblacional es el propuesto.

En el ejemplo, puesto que el nivel crítico es, en ambos casos, menor que 0,05, deben rechazarse ambas hipótesis nulas: en ninguno de los dos casos es razonable atribuir la diferencia observada al azar. Puede concluirse, por tanto, que el nivel educativo medio de ambas poblaciones es distinto de 13,5. Aunque ambos contrastes son bilaterales (pues en ningún caso se ha hecho explícita ninguna expectativa sobre la dirección en que H_0 podría ser falsa), el valor de las medias muestrales (ver Tabla 9.7) permite concretar que el nivel educativo medio es mayor que 13,5 en la población de hombres y menor que 13,5 en la de mujeres.

Las siguiente columna de la Tabla 9.8 ofrece la diferencia entre las medias muestrales y el valor de prueba ($14{,}43 - 13{,}5 = 0{,}93$ en los hombres y $12{,}37 - 13{,}5 = -1{,}13$ en las mujeres). Estas diferencias son el numerador del estadístico T.

Y a continuación aparecen los límites inferior y superior de los intervalos de confianza (calculados al 95%) para las diferencias de la columna anterior. Los límites de estos intervalos se calculan sumando y restando a la diferencia entre cada media muestral y el valor de prueba una cantidad que se obtiene multiplicando el error típico de la media ($\hat{\sigma}_{\bar{Y}} = S_Y/\sqrt{n}$) por el cuantil $100(1-\alpha/2) = 100(1-0{,}05/2) = 97{,}5$ de la distribución t de *Student* con $n-1$ grados de libertad (α se refiere al nivel de significación adoptado). Debe tenerse en cuenta que el intervalo de confianza no se construye a partir de la media muestral, como se hace en [7.11], sino a partir de la diferencia entre la media muestral y el valor poblacional propuesto como valor de prueba. El intervalo de confianza también permite tomar una decisión sobre la hipótesis nula planteada: si los límites del intervalo no incluyen el valor cero (como ocurre en los dos casos del ejemplo), puede concluirse que los datos muestrales son incompatibles con el valor poblacional propuesto y, en consecuencia, debe rechazarse H_0; si los límites no incluyen el valor cero, no puede rechazarse.

Tabla 9.9. Resumen del procedimiento *Prueba T para una muestra*

Nivel educativo

	Valor de prueba = 13.5					
					95% Intervalo de confianza para la diferencia	
Sexo	t	gl	Sig. (bilateral)	Diferencia de medias	Inferior	Superior
Hombre	5,015	257	,000	,930	,56	1,30
Mujer	-7,159	215	,000	-1,130	-1,44	-,82

Nada se ha dicho sobre el supuesto de normalidad en que se basa la prueba T, pero ya se ha señalado que este supuesto sólo es importante con muestras pequeñas (no obstante, el SPSS permite contrastar este supuesto con el procedimiento **Explorar**).

[15] El nivel crítico nunca vale cero. Sólo un estadístico infinitamente grande (o infinitamente pequeño) tiene un nivel crítico igual a cero. Cuando el SPSS informa de que el nivel crítico vale 0,000 es porque sólo muestra los tres primeros decimales. En estos casos, basta con indicar que el nivel crítico es menor que 0,0005 (p < 0,0005), pues si el tercer decimal vale cero y está redondeado, entonces el cuarto necesariamente es menor que 5.

Apéndice 9

Relación entre la distribución *t*, la distribución χ^2 y la varianza

Ya hemos visto (ver [9.10]) que, si la variable Y se distribuye normalmente con media y desviación típica conocidas, la transformación

$$Z = \frac{\bar{Y} - \mu_Y}{\sigma_{\bar{Y}}} = \frac{\bar{Y} - \mu_Y}{\sigma_Y / \sqrt{n}}$$

se distribuye normalmente con media 0 y desviación típica 1, es decir, $N(0,1)$. También hemos visto (ver [9.11]) que, si se desconoce el valor de la varianza poblacional y se estima mediante la varianza muestral, la transformación

$$T = \frac{\bar{Y} - \mu_Y}{\hat{\sigma}_Y / \sqrt{n}} = \frac{\bar{Y} - \mu_Y}{S_Y / \sqrt{n}}$$

ya no se ajusta a la distribución normal $N(0,1)$ sino a la distribución t con $n-1$ grados de libertad (una distribución muy parecida a la normal tipificada, pero no idéntica; ver Apéndice 5).

El numerador de Z y T es el mismo: $\bar{Y} - \mu$; es decir, una variable aleatoria cuyo valor depende del que tome el estadístico $\bar{Y}$ en la muestra concreta elegida. Pero en el denominador de Z y T hay una diferencia importante: Z se obtiene a partir de σ_Y y T a partir de S_Y. Mientras que el denominador de la transformación Z es una *constante* (un parámetro poblacional cuyo valor es independiente de la muestra elegida), el de la transformación T es una *variable* (un estadístico cuyo valor depende, al igual que el de $\bar{Y}$, de la muestra concreta elegida). La consecuencia de esta diferencia es que, mientras un mismo valor de $\bar{Y}$ siempre ofrece el mismo valor Z independientemente de la muestra elegida, un mismo valor $\bar{Y}$ puede ofrecer diferentes valores T dependiendo de la muestra elegida. De ahí que las distribuciones de Z y de T sean distintas.

La responsable de la diferencia entre las transformaciones Z y T y sus respectivas distribuciones es, claro está, la varianza muestral. Por tanto, conocer las características de la distribución muestral de la varianza (ver Apéndice 6) puede ayudarnos a comprender mejor esa diferencia.

Relación entre la distribución *t* y la varianza

Sabemos por [6.15] que S_Y^2 es un estimador insesgado de σ_Y^2 (el concepto de estimador insesgado se ha estudiado en el apartado *Propiedades de un buen estimador* del Capítulo 7):

$$E(S_Y^2) = \sigma_Y^2$$

Es decir, sabemos que el centro (el valor esperado o media) de la distribución muestral de la varianza es la varianza poblacional. Pero esta deseable propiedad no lo es todo. De la ecuación [6.11] se desprende que la forma de la distribución muestral de la varianza adolece de asimetría positiva, especialmente con muestras pequeñas (ver, en el Apéndice 5, el apartado *La distribución* χ^2, particularmente las Figuras 5.7 a la 5.9). Esto significa que, al elegir muchas muestras de una determinada población, el valor promedio de S_Y^2 será igual a σ_Y^2; pero también significa que, al seleccionar una única muestra, los valores menores que σ_Y^2 son más probables que los valores mayores que σ_Y^2, es decir,

$$P(S_Y^2 < \sigma_Y^2) > P(S_Y^2 > \sigma_Y^2) \quad [9.12]$$

Esto implica que, al calcular el estadístico T utilizando S_Y y el estadístico Z utilizando σ_Y, es más probable encontrar $T > Z$ que $T < Z$.

Ahora bien, la forma de la distribución χ^2 depende de los grados de libertad (ver Apéndice 5, Figura 5.9): conforme éstos van aumentando, la distribución χ^2 y, consecuentemente, la distribución muestral de la varianza, se va volviendo más y más simétrica. Con n tendiendo a infinito se verifica que

$$P(S_Y^2 < \sigma_Y^2) \quad \rightarrow \quad P(S_Y^2 > \sigma_Y^2) \qquad [9.13]$$

Por tanto, cuando los grados de libertad van aumentando, la tendencia de S_Y^2 a infraestimar σ_Y^2 va desapareciendo. Pero además, y esto es lo realmente interesante, cuando los grados de libertad van aumentando, la varianza de S_Y^2 se va haciendo más y más pequeña. Recordemos que la varianza del estadístico S_Y^2 viene dada por (ver ecuación [6.15]):

$$V(S_Y^2) \; = \; 2\sigma_Y^4/(n-1)$$

Puesto que los grados de libertad están dividiendo, cuando n va aumentando, la varianza de S_Y^2 se va haciendo más pequeña; con n tendiendo a infinito, S_Y^2 tiende a σ_Y^2. Consecuentemente, cuando los grados de libertad van aumentando, T tiende a igualarse con Z.

Relación entre las distribuciones t y χ^2

Con lo que ya sabemos es posible comprobar sin excesiva complicación que las distribuciones t y χ^2 están relacionadas y que, efectivamente, la transformación T se distribuye según t con $n-1$ grados de libertad. Una forma de definir la distribución t con $n-1$ grados de libertad es a partir de la transformación

$$t_{n-1} \; = \; \frac{Z}{\sqrt{\chi^2_{n-1}/(n-1)}} \qquad [9.14]$$

donde Z es una variable $N(0, 1)$, χ^2_{n-1} es una variable *ji*-cuadrado con $n-1$ grados de libertad, y ambas variables se asumen independientes entre sí (los grados de libertad de la distribución t provienen de los de la distribución χ^2).

Recordemos que el punto de partida siempre es una variable Y distribuida normalmente. El supuesto de normalidad es necesario, por un lado, para que el numerador de [9.14] se distribuya normalmente; también es necesario para que el denominador se distribuya según *ji*-cuadrado (recordemos que una variable *ji*-cuadrado es la suma de variables Z normalmente distribuidas elevadas al cuadrado); y también es necesario para garantizar que el numerador y el denominador de [9.14] sean independientes (pues la media y la varianza de una distribución normal son independientes; cosa que no ocurre en el resto de las distribuciones teóricas más utilizadas).

Sustituyendo en [9.14] el numerador Z por su valor en [9.10] y el denominador χ^2_{n-1} por su valor en [6.11] se obtiene

$$t_{n-1} \; = \; \frac{\dfrac{\bar{Y} - \mu_Y}{\sigma_Y/\sqrt{n}}}{\sqrt{\dfrac{(n-1)\,S_Y^2/\sigma_Y^2}{(n-1)}}} \; = \; \frac{\bar{Y} - \mu_Y}{S_Y/\sqrt{n}} \qquad [9.15]$$

Lo cual significa que, efectivamente, la transformación T propuesta en [9.11] es una t de *Student* con $n-1$ grados de libertad.

Supuestos del estadístico X^2 de Pearson

Para que la aproximación del estadístico X^2 de Pearson (un valor empírico, muestral) se aproxime de forma aceptable a la distribución χ^2 (una distribución teórica) es necesario que se den algunas condiciones que tienen que ver, básicamente, con la *independencia* de las observaciones y con el *tamaño de las frecuencias esperadas*.

Independencia

Este supuesto se refiere a que las observaciones deben ser independientes entre sí, lo cual significa que la clasificación de una observación en una de las I categorías de la variable no debe afectar a la clasificación de ninguna otra observación ni verse afectada por ella. Al clasificar, por ejemplo, a dos sujetos en las categorías de la variable *tabaquismo*, el hecho de que un sujeto sea clasificado como fumador o no fumador no debe afectar a la clasificación del otro sujeto.

No es infrecuente encontrar incumplimientos de este supuesto. Un ejemplo típico se da cuando se realizan repetidas observaciones de los mismos sujetos, de manera que el número total de observaciones es mayor que el número total de sujetos. Una forma razonable de garantizar la independencia entre las observaciones consiste en hacer que cada observación se corresponda con un sujeto distinto. Pero esta regla no debe acatarse sin más, pues no es del todo segura. Siempre es posible encontrar sujetos distintos que no se comportan de forma independiente; es decir, sujetos distintos que muestran comportamientos similares en las variables que se desea estudiar: miembros de la misma familia, estudiantes de la misma clase, pacientes de un mismo hospital, participantes en un experimento que interactúan entre sí en su actividad cotidiana, miembros de un mismo colectivo social o religioso, etc.

El supuesto de independencia es importante porque es necesario para poder definir la distribución de probabilidad a la que se ajustan las frecuencias de la tabla (ver siguiente apartado).

Tamaño de las frecuencias esperadas

El segundo requisito para que el estadístico de Pearson se ajuste bien a la distribución *ji*-cuadrado tiene que ver con el tamaño de las frecuencias esperadas. Los datos de la Tabla 9.4 se han obtenido seleccionando una muestra aleatoria de tamaño $n = 250$ sujetos y clasificando a cada sujeto como *fumador*, *exfumador* o *no fumador*. Al proceder de esta manera se está utilizando un esquema de muestreo llamado **multinomial**. Recibe este nombre porque las posibles frecuencias resultantes de la clasificación se ajustan a una distribución de probabilidad teórica llamada multinomial (ver Capítulo 3).

Al seleccionar una muestra aleatoria de n casos y clasificarlos en las I categorías exclusivas y exhaustivas de una variable categórica, las frecuencias observadas de la tabla, n_i, constituyen una variable aleatoria (resultado de la clasificación independiente de n observaciones aleatorias) con función de probabilidad:

$$P(n_i) = \frac{n!}{\prod_i n_i!} \prod_i \pi_i^{n_i} \qquad [9.16]$$

donde n_i se refiere a un resultado concreto (es decir, a una combinación particular de frecuencias) y π_i a la probabilidad de que un elemento aleatoriamente seleccionado pertenezca a la categoría i. El símbolo Π (multiplicación) indica que hay que multiplicar los diferentes valores de n_i y de π_i.

Ahora bien, si la distribución conjunta de las I frecuencias n_i sigue el modelo de probabilidad multinomial, la distribución de cada frecuencia por separado sigue el modelo de probabilidad binomial, con parámetros n y π_i (pues seleccionar una observación y clasificarla o no en una categoría concreta es un ensayo de Bernoulli; y el conjunto de observaciones clasificadas en esa categoría es la suma de n_i ensayos independientes de Bernoulli; ver, en el Capítulo 3, el apartado *La distribución binomial*). De ahí

que el valor esperado (frecuencia esperada) de cada categoría venga dado, según hemos visto ya (ver ecuación [9.6]), por

$$E(n_i) = m_i = n\pi_i$$

Pero sabemos que la distribución binomial se aproxima a la distribución normal a medida que n va aumentando (ver, en el Capítulo 5, el apartado *Aproximación de la distribución binomial a la normal*). Y también sabemos que la distribución *ji*-cuadrado es el resultado sumar los cuadrados de puntuaciones Z normalmente distribuidas (ver, en el Apéndice 5, el apartado *Distribución ji-cuadrado*). Estos argumentos nos ponen en la pista de un importante supuesto relacionado con el estadístico de Pearson y su distribución *ji*-cuadrado: se está asumiendo que la distribución de las frecuencias de cada casilla es binomial (cosa que efectivamente es así si las observaciones se clasifican independientemente; de ahí el primero de los supuestos ya discutido) y que, conforme el tamaño muestral va aumentando, la distribución binomial se va aproximando a la normal (cosa que sólo es así si el tamaño de cada casilla es lo bastante grande). Pero el tamaño de cada casilla, como parámetro, es la frecuencia esperada, no la observada (la frecuencia observada es fruto del muestreo, no de las condiciones previas que se establecen para definir las características de cada casilla).

Por tanto, para que el estadístico de Pearson se aproxime efectivamente a la distribución *ji*-cuadrado, es necesario que las frecuencias esperadas sean lo bastante grandes. Pero, ¿qué significa "lo bastante grandes"? Las consideraciones ya hechas en el apartado *Aproximación de la distribución binomial a la normal* del Capítulo 5 pueden servir para responder a esta pregunta. Cuando π_i toma un valor próximo a 0,5, la aproximación es lo bastante buena incluso con tamaños muestrales tan pequeños como 5 e incluso menores. Pero conforme el valor de π_i se va alejando de 0,5, mayor necesita ser el tamaño muestral para que la aproximación de la binomial a la normal resulte satisfactoria.

Estas consideraciones son las que han llevado a Cochran (1952) a proponer, como una especie de guía práctica, que la mayor parte de las frecuencias esperadas (al menos el 80%) sean iguales o mayores que 5 ($n\pi_i \geq 5$). Aunque esta recomendación (recogida en el SPSS) puede llegar a ser demasiado exigente en algunos casos (Bradley y otros, 1979; Camilli y Hopkins, 1978, 1979; Larntz, 1978), es, quizá, la recomendación más conocida y aceptada. Overall (1980), por otro lado, ha señalado que el problema de trabajar frecuencias esperadas pequeñas está, no tanto en la probabilidad asociada al error consistente en rechazar una hipótesis nula verdadera, sino en la asociada al error consistente en no rechazar una hipótesis nula falsa (trataremos esta cuestión en el segundo volumen).

Ejercicios

9.1. Se ha seleccionado una muestra aleatoria de 50 sujetos con fobia a los perros y se les ha aplicado un determinado tratamiento durante dos meses. Se han recuperado por completo 30 sujetos. Considerando que este tipo de síntomas remiten espontáneamente a los dos meses en el 25% de los casos, ¿puede afirmarse que el tratamiento consigue más recuperaciones (R) de las que se dan de forma espontánea? ($\alpha = 0{,}05$).

9.2. Se cree que, en la población de estudiantes universitarios, 1/4 tienen ideología política de derecha, 1/4 de centro y 2/4 de izquierda. Al clasificar una muestra aleatoria de 24 estudiantes se ha obtenido el siguiente resultado: 5 de derecha (D), 8 de centro (C) y 11 de izquierda (I). ¿Son compatibles estos datos con la hipótesis de partida? ($\alpha = 0{,}05$).

9.3. Un terapeuta asegura que dispone de un nuevo tratamiento capaz de recuperar (R) con éxito (sin recaída) al 80% de los toxicómanos. Para contrastar esta afirmación, aplica el tratamiento a una muestra aleatoria de 100 toxicómanos y, tras el periodo de seguimiento, constata que se han producido 27 recaídas. ¿Es compatible este resultado con la afirmación del terapeuta? ($\alpha = 0{,}05$).

9.4. Las puntuaciones del WAIS (Escala de Inteligencia para Adultos de Wechsler) se distribuyen normalmente con media 100. Un psicólogo ha construido una nueva prueba de inteligencia (Y) y desea saber si la media que se obtiene con ella se parece o no a la del WAIS. Para ello, selecciona una muestra aleatoria de 100 sujetos y, tras pasarles la prueba, obtiene una media de 105 y una desviación típica de 16. ¿Qué concluirá el psicólogo con un nivel de confianza de 0,95?

9.5. El resultado del ejercicio anterior indica que la media que se obtiene con la nueva escala de inteligencia es mayor que la media que se obtiene con el WAIS. ¿Entre qué límites puede estimarse que se encuentra la nueva media? ($\alpha = 0{,}05$).

9.6. Algunos datos recogidos durante los últimos años señalan que los trastornos de tipo depresivo (D) afectan al 32% de las personas en paro. Un investigador social sospecha que esta cifra es demasiado alta y decide obtener alguna evidencia sobre ello. Selecciona una muestra aleatoria de 300 sujetos en paro y encuentra que 63 de ellos muestran trastornos de tipo depresivo. Utilizando $\alpha = 0{,}01$, ¿qué puede concluirse sobre la sospecha del investigador?

9.7. Se sabe que el número de nacimientos (n_N) no se distribuye homogéneamente entre los siete días de la semana. Quizá por la incomodidad que supone para el personal sanitario, quizá porque las madres prefieren pasar el fin de semana en casa, quizá por alguna otra razón, lo cierto es que los nacimientos son más frecuentes entre semana que en fin de semana. La siguiente tabla muestra cómo se distribuyen entre los días de la semana los 280 nacimientos registrados durante un año en una determinada localidad. ¿Permiten estos datos afirmar que el número de nacimientos no se distribuye homogéneamente entre los días de la semana? ($\alpha = 0{,}05$).

Días de la semana	*L*	*M*	*X*	*J*	*V*	*S*	*D*
Número de nacimientos	45	50	40	55	35	30	25

9.8. El análisis llevado a cabo en el ejercicio anterior indica que, efectivamente, los nacimientos no se distribuyen homogéneamente a lo largo de los días de la semana. Esta afirmación, sin embargo, es bastante imprecisa. Averiguar, con $\alpha = 0{,}05$, en qué días de la semana la proporción de nacimientos difiere significativamente de la esperada.

9.9. La información que ofrece el editor de una escala de madurez señala que las puntuaciones en la escala se distribuyen normalmente con media 5 en la población de estudiantes de enseñanza primaria. La escala tiene ya 10 años, lo que hace sospechar a un educador que el promedio de la escala ha podido aumentar. Para comprobarlo, selecciona una muestra aleatoria de 25 estudiantes de enseñanza primaria y, tras pasarles la prueba, obtiene una media de 5,6 y una desviación típica de 2. ¿Podrá el educador concluir, con $\alpha = 0{,}05$, que el promedio de la escala de madurez ha aumentado?

9.10. Dos psiquiatras han evaluado a 10 pacientes con el propósito de diagnosticar cuáles de ellos padecen pseudoalucinaciones. El informe de los psiquiatras incluye únicamente un *sí* o un *no* para indicar la presencia o ausencia de pseudoalucinaciones. Los datos de los informes de ambos psiquiatras están resumidos en la siguiente tabla. ¿Puede afirmarse que el grado de acuerdo que han alcanzado los psiquiatras es mayor que el que cabría esperar por azar? ($\alpha = 0{,}05$).

Pacientes	1	2	3	4	5	6	7	8	9	10
Psiquiatra 1	Sí	Sí	No	Sí	No	No	Sí	Sí	No	Sí
Psiquiatra 2	Sí	No	Sí	Sí	No	No	Sí	Sí	No	No

9.11. Con el método que se viene utilizado en los últimos años en un determinado colegio para enseñar matemáticas, los estudiantes de enseñanza primaria vienen consiguiendo, en matemáticas, una calificación media de 6,4. Un educador convence al director del centro de que existe un método más eficaz y éste decide aplicarlo durante un año en dos aulas de 25 estudiantes cada una. Al final del curso, los 50 estudiantes obtienen una calificación media de 6,8 y una varianza de 2. ¿Se puede concluir, con un nivel de confianza de 0,95, que el nuevo método de enseñanza ha mejorado la calificación media que se venía obteniendo con el método tradicional?

9.12. En una muestra aleatoria de 10 sujetos con problemas de enuresis se ha aplicado un tratamiento cognitivo-conductual y se han obtenido resultados positivos en 7 casos. ¿Es compatible este resultado con la hipótesis de que al menos el 90% de los sujetos enuréticos podrá recuperarse (R) con este tratamiento? ($\alpha = 0{,}05$).

9.13. ¿A qué conclusión se habría llegado en el ejercicio anterior si, utilizando el mismo contraste y el mismo nivel de significación, en lugar de obtener 7 recuperaciones de 10, se hubieran obtenido 14 recuperaciones de 20 (es decir, si se hubiera obtenido la misma proporción de recuperaciones pero con una muestra mayor? ¿A qué se debe la diferencia en la conclusión?

9.14. Para contrastar la hipótesis de que el 70% de los estudiantes de psicología son mujeres se extrae aleatoriamente de esa población una muestra de 50 estudiantes. ¿Qué número de mujeres (n_M) debe encontrarse para no rechazar la hipótesis $\pi_M = 0{,}70$? ($\alpha = 0{,}05$).

9.15. Teniendo en cuenta que en la muestra de 50 estudiantes de psicología del ejercicio anterior había 39 mujeres, ¿entre qué límites cabe estimar que se encuentra la proporción de mujeres en la población de estudiantes de psicología? ($\alpha = 0{,}05$).

9.16. Un profesor ha diseñado una prueba de aptitud con 17 preguntas dicotómicas. ¿Qué número mínimo de aciertos (A) debe tener un sujeto para poder afirmar, con $\alpha = 0{,}01$, que no ha respondido al azar?

9.17. En un contraste bilateral de H_0: $\mu_Y = 420$, ¿qué valores de $\overline{Y}$ llevarán a rechazar H_0 con una muestra aleatoria de tamaño 36 extraída de una población normal cuya desviación típica vale 18? ($\alpha = 0{,}05$).

9.18. Se ha utilizado el estadístico X^2 de Pearson para contrastar la hipótesis $H_0: f(n_1) = B(n, \pi_1)$. En una muestra aleatoria se ha obtenido $X^2 = 11{,}41$. Sabiendo que $P(X^2 > 11{,}41) = 0{,}185$:

a. ¿Qué decisión debe tomarse sobre H_0, con $\alpha = 0{,}05$?
b. ¿Por qué?
c. ¿Qué puede concluirse?

9.19. Un sociólogo está interesado en averiguar si el *salario* depende del *nivel educativo*. Para ello, selecciona una muestra aleatoria de 900 sujetos y registra, para cada uno, su nivel educativo (sin estudios, primarios, secundarios, medios, superiores) y su salario anual (hasta 20.000 euros, entre 20.000 y 40.000, entre 40.000 y 60.000, entre 60.000 y 80.000 y más de 80.000). Tras analizar los datos con la prueba X^2 de Pearson, toma la decisión de rechazar la hipótesis nula de independencia ($p = 0{,}001$). En este escenario, ¿cuál de las siguientes afirmaciones es verdadera y cuál falsa?

a. El salario depende del nivel educativo.
b. El salario no depende del nivel educativo.
c. El salario no está relacionado con el nivel educativo.
d. Ha quedado probado de forma inequívoca que el salario está relacionado con el nivel educativo.
e. No existe evidencia suficiente para poder afirmar que el salario está relacionado con el nivel educativo.
f. Existe un riesgo de 0,001 de que el investigador haya tomado una decisión equivocada.

9.20. Consideremos la hipótesis $H_0: f(X) = \text{Multinomial } (n = 200;\ \pi_1 = 0{,}50,\ \pi_2 = 0{,}30,\ \pi_3 = 0{,}20)$ y la siguiente tabla:

	X_1	X_2	X_3
n_i	( )	( )	( )
m_i	( )	( )	( )

Sabiendo que, tras recoger los datos y analizarlos, se ha obtenido para el estadístico X^2 de Pearson un valor de cero:

a. Completar la tabla.
b. ¿Qué decisión debe tomarse sobre H_0?
c. ¿Cuánto vale el nivel crítico (valor p) del contraste?

9.21. A continuación se ofrece una tabla con la función de distribución de una variable aleatoria (n_1) distribuida $B(n = 5;\ \pi_1 = 0{,}25)$:

n_1	0	1	2	3	4	5
$F(n_1)$	0,237	0,633	0,896	0,984	0,999	1,000

Si un sujeto responde a un test de 5 preguntas cada una de las cuales tiene 4 alternativas de respuesta de las que sólo una es correcta, ¿cuántas preguntas tiene que acertar *como mínimo* para que se pueda rechazar la hipótesis de que ha respondido al azar? ($\alpha = 0{,}05$).

9.22. En el ejercicio 9.6 se ha estudiado la proporción de personas afectadas de trastorno de tipo depresivo en la población de desempleados. En ese estudio, se ha definido una población (personas en paro), se ha seleccionado una muestra aleatoria, se ha calculado un estadístico (proporción) tras definir el objeto de estudio (trastorno de tipo depresivo) y se han efectuado las inferencias pertinentes. Como consecuencia de todo esto, sabemos que la proporción de desempleados que padecen trastorno depresivo es menor que 0,32. En este escenario, ¿cuál de las siguientes afirmaciones es verdadera y cuál falsa?

a. El desempleo está relacionado con los trastornos depresivos.
b. El desempleo no está relacionado con los trastornos depresivos.
c. El desempleo hace aumentar la proporción de personas con trastorno depresivo.
d. El desempleo hace disminuir la proporción de personas con trastorno depresivo.
e. El desempleo ha disminuido y ésa es la razón por la cual ha disminuido también la proporción de personas con trastornos depresivos.

Soluciones

9.1. Una variable dicotómica (recuperarse, no recuperarse).
Contraste sobre una proporción.

1. *Hipótesis*: H_0: $\pi_R \leq 0{,}25$; H_1: $\pi_R > 0{,}25$ (contraste unilateral derecho).
2. *Supuestos*: muestra aleatoria de tamaño 50 extraída de una población dicotómica con probabilidad de *recuperación* constante en cada extracción.
3. *Estadístico del contraste* (con $n = 50$, $\pi_R = 0{,}25$ y $n_R = 30$):

$$Z = [n_R - n\pi_R]/\sqrt{n\pi_R(1-\pi_R)} = [30 - 50(0{,}25)]/\sqrt{50(0{,}25)(1-0{,}25)} = 5{,}72.$$

4. *Distribución muestral*: Z se distribuye $N(0, 1)$.
5. *Zona crítica*: $Z \geq Z_{0,95} = 1{,}645$.
6. *Decisión*: el valor del estadístico del contraste (5,72) es mayor que el punto crítico (1,645); por tanto, se rechaza H_0. Puede afirmarse que la proporción de recuperaciones que consigue la terapia es mayor que la esperable por recuperación espontánea.

9.2. Una variable categórica (ideología política, con tres categorías).
Contraste sobre bondad de ajuste.

1. *Hipótesis*: $H_0: \pi_D = 1/4$; $\pi_C = 1/4$; $\pi_I = 2/4$.
2. *Supuestos*: muestra aleatoria de $n = 24$ observaciones clasificada en tres categorías con probabilidad constante.
3. *Estadístico del contraste* ($m_D = 24\,(1/4) = 6$; $m_C = 24\,(1/4) = 6$; $m_I = 24\,(2/4) = 12$):

$$X^2 = \Sigma(n_i - m_i)^2/m_i = (5-6)^2/6 + (8-6)^2/6 + (11-12)^2/12 = 0{,}92.$$

4. *Distribución muestral*: X^2 se distribuye según $\chi^2_{I-1} = \chi^2_2$.
5. *Zona crítica*: $X^2 \geq \chi^2_{2;\,0,95} = 5{,}99$.
6. *Decisión*: el valor del estadístico del contraste (0,92) es menor que el punto crítico (5,99); por tanto, se mantiene H_0. Puede concluirse que los datos son compatibles con la hipótesis planteada.

9.3. Una variable dicotómica (recuperarse, no recuperarse).
Contraste sobre una proporción.

1. *Hipótesis*: $H_0: \pi_R = 0{,}80$; H_1: $\pi_R \neq 0{,}80$ (contraste bilateral).

2. *Supuestos*: muestra aleatoria de tamaño 100 extraída de una población dicotómica con probabilidad de *recuperación* constante en cada extracción.
3. *Estadístico del contraste* (con $n = 100$, $\pi_R = 0{,}80$ y $n_R = 100 - 27 = 73$):

$$Z = [n_R - n\pi_R]/\sqrt{n\pi_R(1-\pi_R)} = [73 - 100(0{,}80)]/\sqrt{100(0{,}80)} = -1{,}75.$$

4. *Distribución muestral*: Z se distribuye $N(0, 1)$.
5. *Zona crítica*: $Z \leq Z_{0,025} = -1{,}96$ y $Z \geq Z_{0,975} = 1{,}96$.
6. *Decisión*: el valor del estadístico del contraste ($-1{,}75$) se encuentra entre los puntos críticos $-1{,}96$ y 1,96. Por tanto, no puede rechazarse H_0. Puede concluirse que el resultado obtenido es compatible con la afirmación del terapeuta.

9.4. Una variable cuantitativa (Y = puntuaciones en la nueva escala de inteligencia).
Contraste sobre una media con σ_Y desconocida.

1. *Hipótesis*: H_0: $\mu_Y = 100$; H_1: $\mu_Y \neq 100$ (contraste bilateral).
2. *Supuestos*: muestra aleatoria extraída de una población normal con σ_Y desconocida.
3. *Estadístico del contraste*: $T = (\bar{Y} - \mu_Y)/(S_Y/\sqrt{n}) = (105 - 100)/(16/\sqrt{100}) = 3{,}13$.
4. *Distribución muestral*: T se distribuye según $t_{n-1} = t_{99}$.
5. *Zona crítica*: $T \leq t_{99;\,0,025} = -1{,}984$ y $T \geq t_{99;\,0,975} = 1{,}984$.
6. *Decisión*: como el valor del estadístico del contraste (3,13) es mayor que el punto crítico (1,984), se rechaza H_0. Por tanto, se puede concluir que la media de la nueva escala de inteligencia es distinta de la del WAIS.

9.5. Intervalo de confianza para la media (ecuación [7.11]), con $S_Y/\sqrt{n} = 16/\sqrt{100} = 1{,}6$:

$$IC_{\mu_Y} = \bar{Y} \pm t_{99;\,0,975}\, S_Y/\sqrt{n} = 105 \pm 1{,}984\,(1{,}6) = 105 \pm 3{,}17 = (101{,}83;\ 108{,}17).$$

Podemos estimar, con una confianza del 95%, que la media de la nueva escala se encuentra entre 101,83 y 108,17.

9.6. Una variable dicotómica (padecer o no trastorno depresivo).
Contraste sobre una proporción.

1. *Hipótesis*: H_0: $\pi_D \geq 0{,}32$; H_1: $\pi_D < 0{,}32$ (contraste unilateral izquierdo).
2. *Supuestos*: muestra aleatoria de tamaño 300 extraída de una población dicotómica, con probabilidad de *padecer trastorno depresivo* constante en cada extracción.
3. *Estadístico del contraste* (con $n = 300$, $\pi_D = 0{,}32$ y $n_D = 63$):

$$Z = [n_D - n\pi_D]/\sqrt{n\pi_D(1-\pi_D)} = [63 - 300(0{,}32)]/\sqrt{300(0{,}32)(1-0{,}32)} = -4{,}08.$$

4. *Distribución muestral*: Z se distribuye $N(0, 1)$.
5. *Zona crítica*: $Z \leq Z_{0,05} = -2{,}33$.
6. *Decisión*: como el valor del estadístico del contraste ($-4{,}08$) es menor que el punto crítico ($-2{,}33$), se rechaza H_0. Por tanto, puede concluirse que, en la población de desempleados, la proporción de personas con trastornos de tipo depresivo es menor que 0,32.

9.7. Una variable categórica (días de la semana).
Contraste sobre bondad de ajuste.

1. *Hipótesis*: H_0: $\pi_L = \pi_M = \pi_X = \pi_J = \pi_V = \pi_D = 1/7$.
2. *Supuestos*: muestra aleatoria de $n = 280$ observaciones clasificadas independientemente en las siete categorías de una variable.
3. *Estadístico del contraste* ($m_L = m_M = m_X = m_J = m_V = m_S = m_D = 280\,(1/7) = 40$):

$$X^2 = \Sigma(n_i - m_i)^2/m_i = (45-40)^2/40 + (50-40)^2/40 + \cdots + (25-40)^2/40 = 17{,}5.$$

4. *Distribución muestral*: X^2 se distribuye según $\chi^2_{I-1} = \chi^2_6$.

5. *Zona crítica*: $X^2 \geq \chi^2_{6;\,0,95} = 12{,}59$.
6. *Decisión*: como el valor del estadístico del contraste (17,5) es mayor que el punto crítico (12,59), se rechaza H_0. Puede concluirse que los nacimientos no se reparten homogéneamente entre los días de la semana.

9.8. Intervalos de confianza sobre proporciones.
La siguiente tabla muestra los límites inferior y superior del intervalo de confianza construido para la proporción de nacimientos en cada categoría de la variable *días de la semana*. Estos intervalos se han calculado aplicando la ecuación [9.8]:

$$IC_{\pi_L} = 0{,}161 \pm 1{,}96\sqrt{0{,}161\,(1-0{,}161)/280} = 0{,}161 \pm 0{,}043 = (0{,}118;\ 0{,}204).$$

$$IC_{\pi_M} = 0{,}179 \pm 1{,}96\sqrt{0{,}179\,(1-0{,}179)/280} = 0{,}179 \pm 0{,}045 = (0{,}134;\ 0{,}224).$$

$$IC_{\pi_X} = 0{,}143 \pm 1{,}96\sqrt{0{,}143\,(1-0{,}143)/280} = 0{,}143 \pm 0{,}041 = (0{,}102;\ 0{,}184).$$

$$IC_{\pi_J} = 0{,}196 \pm 1{,}96\sqrt{0{,}196\,(1-0{,}196)/280} = 0{,}196 \pm 0{,}046 = (0{,}150;\ 0{,}242).$$

$$IC_{\pi_V} = 0{,}125 \pm 1{,}96\sqrt{0{,}125\,(1-0{,}125)/280} = 0{,}125 \pm 0{,}039 = (0{,}086;\ 0{,}164).$$

$$IC_{\pi_S} = 0{,}107 \pm 1{,}96\sqrt{0{,}107\,(1-0{,}107)/280} = 0{,}107 \pm 0{,}036 = (0{,}071;\ 0{,}143).$$

$$IC_{\pi_D} = 0{,}089 \pm 1{,}96\sqrt{0{,}089\,(1-0{,}089)/280} = 0{,}089 \pm 0{,}033 = (0{,}056;\ 0{,}122).$$

Días semana	n_i	$P_i = n_i / n$	L_i	L_s	$\pi_i = 1/7 = 0{,}143$
Lunes	45	45/280 = 0,161	0,118	0,204	incluida en el intervalo
Martes	50	50/280 = 0,179	0,134	0,224	incluida en el intervalo
Miércoles	40	40/280 = 0,143	0,102	0,184	incluida en el intervalo
Jueves	55	55/280 = 0,196	0,150	0,242	no incluida en el intervalo
Viernes	35	35/280 = 0,125	0,086	0,164	incluida en el intervalo
Sábado	30	30/280 = 0,107	0,071	0,143	incluida en el intervalo
Domingo	25	25/280 = 0,089	0,056	0,122	no incluida en el intervalo

Las proporciones observadas que los intervalos de confianza delatan como significativamente distintas de la proporción esperada (0,143) son las correspondientes al jueves (donde hay más nacimientos de lo esperado) y al domingo (donde hay menos nacimientos de lo esperado).

9.9. Una variable cuantitativa (Y = puntuaciones en la escala de madurez).
Contraste sobre una media con σ_Y desconocida.
1. *Hipótesis*: H_0: $\mu_Y \leq 5$; H_1: $\mu_Y > 5$ (contraste unilateral derecho).
2. *Supuestos*: muestra aleatoria extraída de una población normal con σ_Y desconocida.
3. *Estadístico del contraste*: $T = (\bar{Y} - \mu_Y)/(S_Y/\sqrt{n}) = (5{,}6-5)/(2/\sqrt{25}) = 1{,}5$.
4. *Distribución muestral*: T se distribuye según $t_{n-1} = t_{24}$.
5. *Zona crítica*: $T \geq t_{24;\,0,95} = 1{,}711$.
6. *Decisión*: como el valor del estadístico del contraste (1,5) es menor que el punto crítico (1,711), se mantiene H_0. Por tanto, los datos obtenidos no permiten afirmar que la media de la escala de madurez haya aumentado.

9.10. Si los diagnósticos se hacen al azar, la probabilidad de que se dé una coincidencia vale 0,50 pues, de cuatro posibilidades (sí-sí, sí-no, no-sí, no-no), hay dos en las que los diagnósticos coinciden. Por

tanto, se trata de averiguar si la proporción de coincidencias (π_C) en el diagnóstico de los dos psiquiatras es mayor que 0,50.

1. *Hipótesis*: H_0: $\pi_R \leq 0{,}50$; H_1 : $\pi_R > 0{,}50$ (contraste unilateral derecho).
2. *Supuestos*: muestra aleatoria de tamaño 10 extraída de una población dicotómica, con probabilidad de recuperación constante en cada extracción.
3. *Estadístico del contraste* (con $n = 10$, $\pi_C = 0{,}50$ y $n_C = 7$).
4. *Distribución muestral*: n_C se distribuye $B(10;\ 0{,}50)$.
5. *Nivel crítico*: $p = P(n_C \geq 7) = 1 - F(6) = 1 - 0{,}828 = 0{,}172$.
6. *Decisión*: dado que el nivel crítico ($p = 0{,}172$) es mayor que el nivel de significación establecido ($\alpha = 0{,}05$), no es posible rechazar H_0. Por tanto, puede concluirse que el grado de acuerdo que han alcanzado los psiquiatras no supera el esperable por azar.

9.11. Una variable cuantitativa (Y = calificaciones en matemáticas).
Contraste sobre una media con σ_Y desconocida.

1. *Hipótesis*: H_0: $\mu_Y \leq 6{,}4$; H_1: $\mu_Y > 6{,}4$ (contraste unilateral derecho).
2. *Supuestos*: muestra aleatoria extraída de una población normal con σ_Y desconocida.
3. *Estadístico del contraste*: $T = (\bar{Y} - \mu_Y)/(S_Y/\sqrt{n}) = (6{,}8 - 6{,}4)/(\sqrt{2/50}) = 2$.
4. *Distribución muestral*: T se distribuye según $t_{n-1} = t_{49}$.
5. *Zona crítica*: $T \geq t_{49;\,0{,}95} \approx 1{,}676$.
6. *Decisión*: como el valor del estadístico del contraste (2) es mayor que el punto crítico (1,676), se rechaza H_0. Por tanto, se puede concluir que el rendimiento medio que se obtiene con el nuevo método es mayor que el que se venía obteniendo con el método tradicional.

9.12. Una variable dicotómica (recuperarse, no recuperarse).
Contraste sobre una proporción.

1. *Hipótesis*: H_0: $\pi_R \geq 0{,}90$; H_1 : $\pi_R < 0{,}90$ (contraste unilateral izquierdo).
2. *Supuestos*: muestra aleatoria de tamaño 10 extraída de una población dicotómica, con probabilidad de recuperación constante en cada extracción.
3. *Estadístico del contraste*: $n_R = 7$.
4. *Distribución muestral*: n_R se distribuye $B(10;\ 0{,}90)$.
5. *Nivel crítico*: $p = P(n_R \leq 7) = 0{,}070$.
6. *Decisión*: dado que el nivel crítico ($p = 0{,}070$) es mayor que el nivel de significación establecido ($\alpha = 0{,}05$), no es posible rechazar H_0. Por tanto, el resultado obtenido es compatible con la hipótesis de que al menos el 90% de los pacientes podrán recuperarse con el tratamiento.

9.13. Una variable dicotómica (recuperarse, no recuperarse).
Contraste sobre una proporción.

1. *Hipótesis*: H_0: $\pi_R \geq 0{,}90$; H_1 : $\pi_R < 0{,}90$ (contraste unilateral izquierdo).
2. *Supuestos*: muestra aleatoria de tamaño 20 extraída de una población dicotómica, con probabilidad de recuperación constante en cada extracción.
3. *Estadístico del contraste*: $n_R = 14$.
4. *Distribución muestral*: n_R se distribuye $B(20;\ 0{,}90)$.
5. *Nivel crítico*: $P(n_R \leq 14) = 0{,}011$.
6. *Decisión*: dado que el nivel crítico ($p = 0{,}011$) es menor que el nivel de significación establecido ($\alpha = 0{,}05$), se rechaza H_0. Por tanto, no puede afirmarse que al menos el 90% de los pacientes podrá recuperarse con el tratamiento.

En ambos casos $P_R = 7/10 = 14/20 = 0{,}70$. Con $n = 20$, la diferencia entre la proporción teórica y la empírica ($\pi_R - P_R = 0{,}90 - 0{,}70 = 0{,}20$) es declarada estadísticamente significativa; con $n = 10$, esa diferencia no es lo bastante grande como para descartar que pueda producirse por azar. La razón de esto es que al aumentar el tamaño muestral disminuye el error típico del estadístico del contraste y eso permite trabajar con mayor precisión.

9.14. En un contraste bilateral sobre una proporción, con $n = 50$ y $\alpha = 0{,}05$, se rechaza la hipótesis nula cuando el estadístico Z toma un valor menor que $Z_{0,025}$ o mayor que $Z_{0,975}$. Por tanto, los valores n_M que llevarán a rechazar H_0 serán los correspondientes a $Z = -1{,}96$ y $Z = 1{,}96$. Estos valores pueden obtenerse a partir de la ecuación [9.2]:

$$Z_{0,025} = -1{,}96 = (n_{M_1} - 50(0{,}70)/\sqrt{(50(0{,}70)(1-0{,}70)} = (n_{M_1} - 35)/3{,}24.$$
$$n_{M_1} = -1{,}96\,(3{,}24) + 35 = 28{,}65.$$

$$Z_{0,975} = 1{,}96 = (n_{M_2} - 50(0{,}70)/\sqrt{(50(0{,}70)(1-0{,}70)} = (n_{M_2} - 35)/3{,}24.$$
$$n_{M_2} = 1{,}96\,(3{,}24) + 35 = 41{,}35.$$

Por tanto, no podrá rechazarse H_0: $\pi_M = 0{,}70$ cuando el número de mujeres en la muestra se encuentre entre 29 y 41.

9.15. Intervalo de confianza para la proporción (ecuación [7.16]), con $P_M = 39/50 = 0{,}78$:

$$IC_{\pi_M} = P_M \pm |Z_{0,025}|\sqrt{P_M(1-P_M)/n} =$$
$$= 0{,}78 \pm 1{,}96\sqrt{0{,}78(1-0{,}78)/50} = 0{,}78 \pm 0{,}11 = (0{,}67;\ 0{,}89).$$

Podemos estimar, con una confianza del 95%, que la proporción de mujeres en la población de estudiantes de psicología se encuentra entre 0,67 y 0,89.

9.16. En el contraste unilateral derecho sobre una proporción, con $n = 17$ y $\alpha = 0{,}01$, se rechaza la hipótesis de que un sujeto está respondiendo al azar ($\pi_A = 0{,}05$) cuando obtiene un número de aciertos (n_A) cuya probabilidad asociada (es decir, la probabilidad de obtener ese número de aciertos o más) es menor que el nivel de significación establecido (0,01). Consultando la tabla de la *distribución binomial* (Tabla B del Apéndice final) se obtiene:

...

$P(n_A \geq 12) = 1 - F(11) = 1 - 0{,}928 = 0{,}072.$
$P(n_A \geq 13) = 1 - F(12) = 1 - 0{,}975 = 0{,}025.$
$P(n_A \geq 14) = 1 - F(13) = 1 - 0{,}994 = 0{,}006.$

...

Por tanto, para poder descartar que un sujeto está respondiendo al azar, debe obtener 14 aciertos o más.

9.17. En el contraste bilateral sobre una media, con σ_Y conocida y $\alpha = 0{,}05$, se rechaza H_0 cuando el estadístico Z toma un valor menor que $Z_{0,025}$ o mayor que $Z_{0,975}$. Por tanto, los valores $\bar{Y}$ que llevarán a rechazar H_0 serán los correspondientes a $Z = -1{,}96$ y $Z = 1{,}96$. Para conocer estos valores basta con aplicar la ecuación [9.10]:

$$Z_{0,025} = -1{,}96 = (\bar{Y}_1 - 420)/(18/\sqrt{36}) \quad \rightarrow \quad \bar{Y}_1 = -1{,}96\,(18/\sqrt{36}) + 420 = 414{,}12.$$
$$Z_{0,975} = 1{,}96 = (\bar{Y}_2 - 420)/(18/\sqrt{36}) \quad \rightarrow \quad \bar{Y}_2 = 1{,}96\,(18/\sqrt{36}) + 420 = 425{,}88.$$

Es decir, rechazaremos H_0 cuando $\bar{Y}$ tome un valor menor que 414,12 o mayor que 425,88.

9.18. *a*. Mantenerla.
b. Porque el nivel crítico asociado al estadístico de contraste ($p = 0{,}185$) es mayor que el nivel de significación adoptado ($\alpha = 0{,}05$). Es decir, porque $P(X^2 > 11{,}41) = 0{,}185 > 0{,}05$.
c. No existe razón para pensar que la variable estudiada no se distribuya según se propone en H_0, es decir, $B(n, \pi_1)$.

9.19. Todas las afirmaciones son falsas. La única conclusión válida de un estudio de estas características es que el *salario está relacionado con el nivel educativo* (esto es lo que implica el rechazo de la hipótesis

de independencia). Esto descarta las alternativas en las que implícita o explícitamente se afirma que no existe relación: *b*, *c* y *e*.

La alternativa *a* afirma que la relación encontrada es de naturaleza causal, y esto es algo que excede el alcance de un estudio no experimental cuando, como es el caso, no existe una teoría previa que justifique el hallazgo.

En la alternativa *d* se hace una afirmación categórica que no se corresponde con la lógica del contraste de hipótesis. Un contraste de hipótesis se basa en probabilidades y, por tanto, no es una prueba definitiva o inequívoca de nada.

El nivel crítico ($p = 0{,}001$) es una probabilidad *condicional* referida a los datos (al estadístico del contraste). Esto no debe confundirse con una probabilidad referida a las hipótesis. Y una afirmación sobre la probabilidad de tomar una decisión equivocada cuando se decide rechazar una hipótesis es una afirmación sobre la probabilidad de que la hipótesis rechazada sea verdadera. Esto descarta la alternativa *f*.

9.20. *a*. Si $X^2 = 0$, entonces todas las frecuencias observadas son iguales que sus correspondientes frecuencias esperadas: $n_i = m_i = n\text{p}_i$. Por tanto:

	X_1	X_2	X_3
n_i	100	60	40
m_i	100	60	40

b. Mantenerla.

c. $p = P(X^2 \geq 0) = 1$.

9.21. La probabilidad de acertar por azar una pregunta con 4 alternativas de respuesta vale $1/4 = 0{,}25$; por tanto, la función de probabilidad que se ofrece la tabla no es más que la distribución muestral del estadístico *número de aciertos por azar*. La probabilidad de obtener por azar $n_1 = 3$ aciertos o más vale $1 - 0{,}896 = 0{,}104$. En consecuencia, con 3 aciertos y habiendo fijado $\alpha = 0{,}05$ como criterio de decisión, no es posible rechazar la hipótesis de que el sujeto ha respondido al azar. La probabilidad de acertar 4 preguntas o más por azar vale $1 - 0{,}984 = 0{,}016$, y esa probabilidad es menor que $\alpha = 0{,}05$. Por tanto, con 4 aciertos o más puede rechazarse la hipótesis de que un sujeto ha respondido al azar.

9.22. En este estudio, el desempleo no es una variable, sino una constante. Por tanto, no tiene sentido hablar de si el desempleo está o no relacionado con algo. Este argumento sirve para declarar falsas todas las alternativas. Pero, además, en las alternativas *c*, *d* y *e* se está haciendo referencia a una relación de tipo causal entre el desempleo y los trastornos depresivos: incluso aunque se hubiera seleccionado una muestra de personas desempleadas y otra de personas empleadas, el tipo de estudio (no experimental) no permitiría hacer este tipo de afirmaciones.

10

Inferencia con dos variables categóricas

En éste y en los dos próximos capítulos se estudian algunas de las herramientas inferenciales disponibles para el análisis de dos variables. En este capítulo se aborda el análisis de *dos variables categóricas* mediante la prueba X^2 de Pearson; en el siguiente, el de *una variable categórica y una cuantitativa* mediante la prueba T de Student para muestras independientes; y en el último, el de *dos variables cuantitativas* con la prueba T de Student para muestras relacionadas y el análisis de correlación lineal simple.

Esta forma de ordenar la exposición no es arbitraria. Recordemos que la naturaleza de las variables (categórica, cuantitativa) condiciona el tipo de herramientas que permiten extraer información útil de los datos. Y recordemos también que basar la clasificación de los procedimientos estadísticos en la naturaleza de las variables contribuye de forma importante a facilitar la elección del procedimiento apropiado en cada caso.

Ya hemos señalado en capítulos anteriores que las herramientas inferenciales sirven para realizar *comparaciones* y estudiar *relaciones*. Cualquier herramienta inferencial, desde la más simple a la más compleja, cubre uno de esos dos objetivos o ambos. Pero ocurre que, cuando hay variables *categóricas* por medio, *comparar y relacionar son la misma cosa* (en concreto, diferencias = relación). Esto no es así cuando todas las variables son *cuantitativas*; en este caso, *comparar y relacionar son cosas distintas* (volveremos sobre esta idea en el Capítulo 12). Por esta razón, de momento, para analizar dos variables categóricas vamos a estudiar un único procedimiento: la prueba X^2 de Pearson; y también vamos a estudiar un único procedimiento para analizar una variable categórica y una cuantitativa: la prueba T de Student para muestras independientes. Sin embargo, para analizar dos variables cuantitativas vamos a estudiar dos procedimientos: la prueba T de Student (para compararlas) y el análisis de correlación lineal (para relacionarlas).

Es importante recordar, una vez más, que todo análisis inferencial debe ir precedido del correspondiente análisis descriptivo. Por tanto, para estudiar dos variables simultáneamente es necesario, previamente, describir cada una de ellas tanto por separado como de forma conjunta. Ocasionalmente, ese análisis descriptivo servirá para cubrir los objetivos del estudio. Pero, incluso cuando esto no sea así, será necesario detenerse en él antes de pasar a la fase inferencial.

Variables categóricas

En las ciencias sociales y de la salud es frecuente encontrarse con variables categóricas. Son variables, recordemos, de las que únicamente es posible obtener una medida de tipo nominal (u ordinal con pocos valores). En una investigación clínica, por ejemplo, se pueden encontrar variables como *padecer o no* una determinada enfermedad o síntoma, o se puede clasificar a los pacientes como *tratados* y *no tratados*, o como *recuperados* y *no recuperados*, etc. En una investigación de tipo social se puede clasificar a los sujetos por la *actitud* que manifiestan hacia un evento particular (favorable, desfavorable o indiferente), o por su *estado civil* (solteros, casados, viudos, divorciados, separados). El *sexo*, la *raza*, la *ideología política*, el *lugar de procedencia*, la *ocupación laboral*, el *resultado* de una tarea (éxito-fracaso), etc., son otros ejemplos de este tipo de variables.

Al trabajar simultáneamente con dos variables categóricas hay que abordar dos tareas básicas: (1) describir el comportamiento conjunto de ambas variables y (2) averiguar si están relacionadas. La primera tarea se lleva a cabo construyendo tablas de contingencias y gráficos de barras agrupadas; la segunda, aplicando la prueba X^2 de Pearson. En el caso de que exista relación es necesario realizar dos tareas adicionales: (1) cuantificarla con alguna medida de asociación y (2) interpretarla a partir de los residuos tipificados.

Tablas de contingencias

Recordemos que, para describir una variable categórica, se utiliza un tipo particular de resumen llamado *tabla de frecuencias* (ver Capítulo 3). Para describir, por ejemplo, la variable *sexo* (hombres, mujeres), o la variable *tabaquismo* (fumadores, exfumadores, no fumadores), se organizan los datos en resúmenes como los que muestra la Tabla 10.1 (los datos provienen de una muestra de 200 universitarios).

Tabla 10.1. Frecuencias de *sexo* (izquierda) y *tabaquismo* (derecha)

Sexo	n_i
Hombres	94
Mujeres	106
Total	200

Tabaquismo	n_i
Fumadores	60
Exfumadores	13
No fumadores	127
Total	200

Para describir simultáneamente dos variables categóricas se hace algo muy parecido: se construye una tabla de frecuencias *conjuntas* combinando las categorías de ambas variables. La Tabla 10.2 es un ejemplo de este tipo de tablas: los mismos 200 sujetos que en la Tabla 10.1 se han clasificado en dos tablas de frecuencias separadas, ahora se han clasificado en una sola tabla. A esta forma de organizar y resumir los datos se le llama **tabla de contingencias**[1].

[1] El término *contingencia* se refiere a la posibilidad de que algo ocurra. En una tabla de contingencias existen tantas posibilidades de que algo ocurra como combinaciones resultan de cruzar las categorías de las variables que definen la tabla. Por tanto, cada casilla de la tabla representa una posibilidad, es decir, una contingencia; de ahí que al conjunto de casillas de la tabla, es decir, al conjunto de contingencias, se le llame *tabla de contingencias*.

En una tabla de estas características, al contenido de las casillas (parte interior de la tabla) se le llama **frecuencias conjuntas** o, simplemente, frecuencias; y a los totales de cada fila y columna (parte exterior de la tabla) se les llama **frecuencias marginales**.

Las *frecuencias marginales* reproducen las frecuencias de cada variable individualmente considerada (se les suele llamar *distribuciones marginales);* estas frecuencias ya contienen toda la información disponible en las respectivas tablas de frecuencias. Las *frecuencias conjuntas* ofrecen información adicional: indican lo que ocurre en cada casilla, es decir, lo que ocurre en cada combinación entre las categorías de las variables. Y, de la misma manera que las frecuencias de una variable individualmente considerada informan del comportamiento de esa variable, las frecuencias conjuntas informan del comportamiento conjunto de ambas variables (volveremos enseguida sobre esta cuestión).

Tabla 10.2. Tabla de contingencias de *sexo* por *tabaquismo*

	Tabaquismo			
Sexo	Fumadores	Exfumadores	No fumadores	*Total*
Hombres	18	7	69	94
Mujeres	42	6	58	106
Total	60	13	127	200

La Tabla 10.2 es un ejemplo de tabla *bidimensional*. Representa el ejemplo más elemental de tabla de contingencias: sólo dos variables. Pero una tabla de contingencias puede ser más compleja. De hecho, pueden construirse tablas de tres, cuatro, o más dimensiones. El límite en el número de dimensiones de una tabla únicamente viene impuesto por el tipo de situación real que se desee representar y por el grado de complejidad que se esté dispuesto a abordar en la interpretación. No obstante, dado que en este apartado estamos tratando el caso de dos variables, nuestra exposición se limitará al caso de tablas bidimensionales.

La Tabla 10.3 muestra la forma general de presentar los datos en una tabla de contingencias bidimensional y la notación que utilizaremos para identificar cada elemento de la tabla. Las I categorías de la variable X definen las filas de la tabla; para identificar cada una de estas categorías (cada fila), se utiliza el subíndice i; por tanto: $i = 1, 2, ..., I$. Las J categorías de la variable Y definen las columnas de la tabla; para identificar cada una de estas categorías (cada columna) se utiliza el subíndice j: por tanto: $j = 1, 2, ..., J$.

Tabla 10.3. Notación utilizada en tablas de contingencias bidimensionales

	Y							
X	1	2	···	j	···	J	n_{i+}
1	n_{11}	n_{12}	···	n_{1j}	···	n_{1J}	n_{1+}
2	n_{21}	n_{22}	···	n_{2j}	···	n_{1J}	n_{2+}
···	···	···	···	···	···	···	···
i	n_{i1}	n_{i2}	···	n_{ij}	···	n_{iJ}	n_{i+}
···	···	···	···	···	···	···	···
I	n_{I1}	n_{I2}	···	n_{Ij}	···	n_{IJ}	n_{I+}
n_{+j}	n_{+1}	n_{+2}	···	n_{+j}	···	n_{+J}	n

n_{ij} = frecuencias conjuntas de X e Y.
n_{i+} = frecuencias marginales de X.
n_{+j} = frecuencias marginales de Y.
n = número total de casos.

El signo “+” se refiere a todos los valores del subíndice al que sustituye; por tanto, cuando sustituye al subíndice i se refiere a todos los valores de i (es decir, 1, 2, ..., I); y cuando sustituye al subíndice j se refiere a todos los valores de j (es decir, 1, 2, ..., J). Así, por ejemplo, en la Tabla 10.2, la frecuencia conjunta n_{12} vale 7; la frecuencia marginal n_{1+} vale 94; y la frecuencia marginal n_{+2} vale 13. Sumando las frecuencias marginales de las filas (n_{i+}), o las de las columnas (n_{+j}), o las conjuntas (n_{ij}), se obtiene idéntico resultado:

$$\sum_i n_{i+} = \sum_j n_{+j} = \sum_i \sum_j n_{ij} = n \qquad [10.1]$$

Por ejemplo, en los datos de la Tabla 10.2 se verifica:

$$94 + 106 = 60 + 13 + 127 = 18 + 7 + 69 + 42 + 6 + 58 = 200$$

La notación utilizada permite identificar abreviadamente las dimensiones de una tabla de contingencias mediante la expresión $I \times J$, donde el número de letras indica el número de variables (dos letras = dos variables) y el valor de las letras representa el número de categorías de cada variable. Así, la Tabla 10.2 es una tabla 2×3 (dos variables, la primera con 2 categorías y la segunda con 3).

Tipos de frecuencias

Los números que aparecen en una tabla de contingencias son *frecuencias*, no puntuaciones; más concretamente, **frecuencias absolutas** (ya sean conjuntas o marginales). El valor 18 de la primera casilla de la Tabla 10.2 ($n_{11} = 18$) indica que, de los 200 sujetos que componen la muestra, 18 han sido clasificados como *hombres fumadores*; el valor 58 de la última casilla de la tabla ($n_{23} = 58$) indica que 58 sujetos han sido clasificados como *mujeres no fumadoras*

Las frecuencias absolutas pueden transformarse fácilmente en **frecuencias porcentuales**. Ahora bien, en una tabla de contingencias bidimensional hay tres tipos de frecuencias porcentuales: (1) los *porcentajes de fila* indican el porcentaje que cada frecuencia conjunta representa respecto de la frecuencia marginal de su fila; (2) los *porcentajes de columna* indican el porcentaje que cada frecuencia conjunta representa respecto de la frecuencia marginal de su columna; y (3) los *porcentajes del total* indican el porcentaje que cada frecuencia conjunta representa respecto del número total de casos de la tabla.

En la primera casilla de la Tabla 10.2, el *porcentaje del total* vale 18(100)/200 = 9,0%; esto significa que el 9,0% de los sujetos de la muestra son hombres fumadores. Estos porcentajes tienen interés descriptivo, pues indican qué porcentaje de casos ha sido clasificado en cada casilla de la tabla; pero, según veremos enseguida, no contienen información útil para interpretar el comportamiento conjunto de ambas variables.

La información realmente útil está en los *porcentajes de fila* y en los *porcentajes de columna*. A estos porcentajes se les llama **distribuciones condicionales**. La Tabla 10.4 muestra los porcentajes de fila correspondientes a las frecuencias absolutas de la Tabla 10.2. Estos porcentajes contienen las distribuciones condicionales de la variable colocada en las columnas. Por tanto, indican cómo se distribuye la variable *tabaquismo* en cada categoría de la variable *sexo* (nótese que los porcentajes de cada distribución condicional suman 100). El porcentaje de fila de la primera casilla (hombres fumadores) vale 18(100)/94 = 19,1%; este valor indica que el 19,1% de los hombres son fumadores. Los porcentajes marginales de las columnas contienen la distribución marginal (la distribución no condicional) de la variable colocada

en las columnas (*tabaquismo*); estos porcentajes indican que hay un 30% de fumadores, un 6,5% de exfumadores y un 63,5% de no fumadores.

La Tabla 10.5 muestra los porcentajes de columna correspondientes a las frecuencias absolutas de la Tabla 10.2. Estos porcentajes contienen las distribuciones condicionales de la variable colocada en las filas. Por tanto, indican cómo se distribuye la variable *sexo* en cada categoría de la variable *tabaquismo* (nótese que los porcentajes de cada distribución condicional suman 100). El porcentaje de columna de la primera casilla (hombres fumadores) vale 18(100)/60 = 30,0%; este valor indica que el 30,0% de los fumadores son hombres. Los porcentajes marginales de las filas contienen la distribución marginal (la distribución no condicional) de la variable colocada en las filas (*sexo*); estos porcentajes indican que hay un 47% de hombres y un 53% de mujeres.

Según veremos enseguida, las comparaciones entre *distribuciones condicionales* constituyen la clave del estudio de la relación en tablas de contingencias.

Tabla 10.4. Porcentajes de *fila* correspondientes a las frecuencias de la Tabla 10.2

	Tabaquismo			
Sexo	Fumadores	Exfumadores	No fumadores	
Hombres	19,1 %	7,4 %	73,4 %	100,0 %
Mujeres	39,6 %	5,7 %	54,7 %	100,0 %
	30,0 %	6,5 %	63,5 %	100,0 %

Tabla 10.5. Porcentajes de *columna* correspondientes a las frecuencias de la Tabla 10.2

	Tabaquismo			
Sexo	Fumadores	Exfumadores	No fumadores	
Hombres	30,0 %	53,8 %	54,3 %	47,0 %
Mujeres	70,0 %	46,2 %	45,7 %	53,0 %
	100,0 %	100,0 %	100,0 %	100,0 %

Gráficos de barras agrupadas

Las frecuencias de una tabla de contingencias (tanto las absolutas como las porcentuales) constituyen la esencia de lo que podríamos llamar la fase descriptiva del análisis de dos variables categóricas. Sin embargo, la fase descriptiva no debe limitarse a presentar los datos de una tabla de contingencias. Ya hemos señalado en varios capítulos anteriores la conveniencia de acompañar los números de una tabla con gráficos que permitan formase una idea rápida acerca de lo que está ocurriendo.

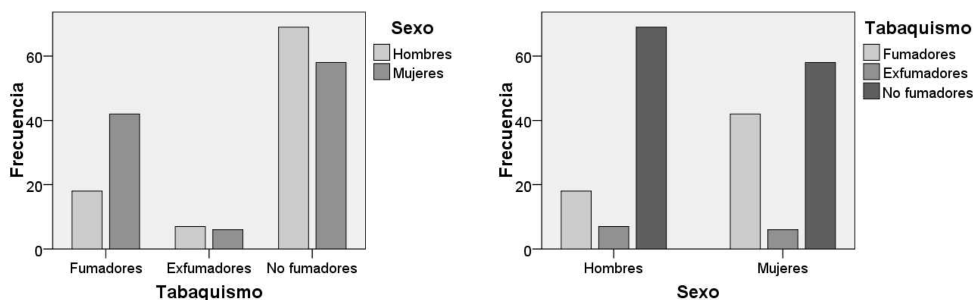
Los gráficos más utilizados para describir simultáneamente dos variables categóricas son los de **barras agrupadas**. Estos gráficos se construyen sobre el plano definido por dos ejes cartesianos: en el eje horizontal se colocan las categorías de una de las variables; en el vertical, las frecuencias conjuntas absolutas; y sobre cada valor de la variable colocada en el eje horizontal se levantan tantas barras como valores tenga la segunda variable. La altura de las

barras es proporcional al tamaño de las frecuencias (la anchura de las barras no es relevante, pero todas ellas han de tener la misma).

La Figura 10.1 muestra el gráfico de barras agrupadas correspondiente a las frecuencias absolutas de la Tabla 10.2. En el gráfico de la izquierda se ha colocado la variable *tabaquismo* en el eje horizontal y se ha levantado una barra por cada categoría de la variable *sexo*; en el de la derecha se ha colocado la variable *sexo* en el eje horizontal y se ha levantado una barra por cada categoría de la variable *tabaquismo*. Ambos gráficos incluyen exactamente el mismo número de barras (tantas como casillas) y exactamente del mismo tamaño (el de las frecuencias absolutas de cada casilla), pero organizadas de distinta manera. La elección entre uno u otro depende únicamente de cuál de los dos permite mostrar con mayor claridad aquello que se quiere resaltar. En ambos gráficos se aprecia que las barras más altas corresponden a los no fumadores (sean hombres o mujeres). Sin embargo, en el de la izquierda destaca el hecho de que la diferencia entre el número de hombres y mujeres es sensiblemente mayor en el grupo de fumadores que en el de no fumadores (la barra clara no alcanza la mitad de la altura de la barra oscura en el grupo de fumadores, pero es más alta que la oscura en el de no fumadores). Y en el gráfico de la derecha destaca el hecho de que la diferencia entre el número de fumadores y no fumadores es sensiblemente mayor entre los hombres que entre las mujeres (la diferencia entre la barra más clara y la más oscura es mayor en el grupo de hombres que en el de mujeres).

Un gráfico de barras agrupadas no se altera si en lugar de representar las *frecuencias absolutas* se representan los *porcentajes referidos al total*. Sin embargo, ya hemos señalado que la información realmente interesante no se encuentra en estos porcentajes, sino en los referidos a las filas y a las columnas, es decir, en las *distribuciones condicionales*. Y las distribuciones condicionales generan gráficos muy diferentes de los que recoge la Figura 10.1 (tanto más diferentes cuanto más distintas entre sí son las frecuencias marginales de la tabla).

Figura 10.1. Gráficos de barras agrupadas basados en las frecuencias absolutas de la Tabla 10.2

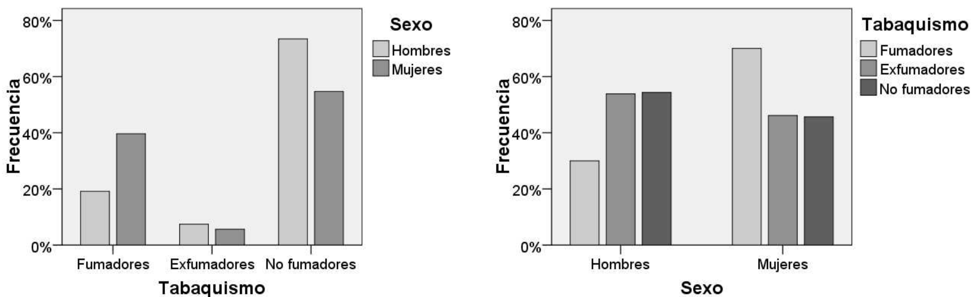


La Figura 10.2 (izquierda) muestra el gráfico correspondiente a los *porcentajes de fila*; por tanto, en él están representadas las distribuciones condicionales de la variable *tabaquismo* (ver Tabla 10.4): las barras claras indican cómo se distribuyen los *fumadores*, los *exfumadores* y los *no fumadores* en el grupo de *hombres*; las oscuras, cómo se distribuyen los *fumadores*, los *exfumadores* y los *no fumadores* en el grupo de mujeres. Como el número de hombres y de mujeres de la muestra es similar (94 y 106), este gráfico se parece bastante al elaborado a partir de las frecuencias absolutas (gráfico de la izquierda en la Figura 10.1). Para interpretarlo correctamente no hay que olvidar que la altura de las barras se basa en los porcentajes referidos a las filas. El dato más llamativo del gráfico es la diferencia existente entre ambas

distribuciones condicionales en lo referente al porcentaje de fumadores y no fumadores: en el grupo de hombres, el porcentaje de fumadores no llega ni a un tercio del de no fumadores; en el de mujeres, el porcentaje de fumadoras es más de la mitad del de no fumadoras.

La Figura 10.2 (derecha) muestra el gráfico de barras correspondiente a los *porcentajes de columna*; por tanto, en él están representadas las distribuciones condicionales de la variable *sexo* (ver Tabla 10.5): las barras más claras indican cómo se distribuyen los *hombres* y las *mujeres* en el grupo de *fumadores*; las menos claras indican cómo se distribuyen los *hombres* y las *mujeres* en el grupo de *exfumadores*; las más oscuras indican cómo se distribuyen los *hombres* y las *mujeres* en el grupo de *no fumadores*. Como el número de fumadores, exfumadores y no fumadores es muy distinto (60,13 y 127), este gráfico es muy diferente del elaborado a partir de las frecuencias absolutas (gráfico de la derecha en la Figura 10.1). Para interpretarlo correctamente no hay que olvidar que la altura de las barras se basa en los porcentajes referidos a las columnas. Quizá el dato más llamativo ahora es que la distribución de la variable sexo (porcentaje de hombres y mujeres) en el grupo de fumadores (barras más claras) sigue una pauta muy distinta de esa misma distribución en los otros dos grupos; en concreto, mientras que en el grupo de fumadores el porcentaje de hombres no llega a la mitad del de mujeres, en los otros dos grupos el porcentaje de hombres es mayor que el de mujeres.

Figura 10.2. Gráficos de barras agrupadas basados en los porcentajes de fila de la Tabla 10.4 (gráfico de la izquierda) y en los porcentajes de columna de la Tabla 10.5 (gráfico de la derecha)



Asociación en tablas de contingencias

Si se combinan dos variables en una tabla de contingencias no es, obviamente, para estudiar cómo se comporta cada una de ellas por separado; para analizarlas individualmente no hace falta combinarlas. La finalidad de combinar dos variables en una sola tabla es la de averiguar si tienen o no algo que ver entre sí, es decir, si están o no relacionadas.

Con dos variables categóricas sólo caben dos posibilidades: o las variables son independientes o están relacionadas[2]. Para aclarar los conceptos de *independencia* y *relación* entre variables categóricas, retomemos los porcentajes de fila de la Tabla 10.4 (basados en las frecuencias absolutas de la Tabla 10.2). De acuerdo con la distribución marginal de la variable

[2] Con tres variables (X, Y, Z) pueden darse diferentes pautas de asociación: puede que las tres variables sean independientes entre sí; puede que sólo haya relación entre X e Y; puede que sólo haya relación entre X y Z; puede que haya relación *entre* X e Y y entre X y Z, pero no entre Y y Z; etc. Sin embargo, con dos variables sólo existen dos posibilidades.

tabaquismo, el 30,0% de los sujetos son fumadores, el 6,5% exfumadores y el 63,5% no fumadores. Pues bien, si la variable *tabaquismo* fuera independiente de la variable *sexo*, entre los hombres también debería ocurrir que un 30,0% son fumadores, un 6,5% son exfumadores y un 63,5% son no fumadores; y lo mismo debería ocurrir entre las mujeres: 30,0%, 6,5% y 63,5% de fumadoras, exfumadoras y no fumadoras, respectivamente. Por tanto, decimos que dos variables son **independientes** *cuando el comportamiento de una de ellas no se ve alterado por la presencia de la otra*. La variable *tabaquismo* será independiente de la variable *sexo* si el comportamiento individual de la variable *tabaquismo* (30,0%, 6,5%, 63,5%) no se ve alterado por la presencia de la variable *sexo*, es decir, si sigue habiendo un 30,0% de fumadores, un 6,5% de exfumadores y un 63,5% de no fumadores tanto en el grupo de hombres como en el de mujeres. Por tanto,

> *Decimos que dos variables categóricas son* ***independientes*** *cuando las distribuciones condicionales de cualquiera de ellas son iguales en todas las categorías de la otra.*

Esto equivale a afirmar que dos variables categóricas son independientes cuando las distribuciones condicionales de ambas son iguales a sus respectivas distribuciones marginales. Cuando no se da esta circunstancia, decimos que las variables están relacionadas o **asociadas**[3]. De manera más formal, decimos que dos sucesos, A y B, son independientes si la probabilidad de su intersección (es decir, la probabilidad de su verificación conjunta o simultánea) es igual al producto de sus probabilidades individuales[4]; es decir, decimos que los sucesos A y B son independientes si $P(A \cap B) = P(A)P(B)$. Trasladando esta definición a los sucesos de una tabla de contingencias, puede afirmarse que el suceso i es independiente del suceso j si:

$$P(X = X_i \cap Y = Y_j) \;=\; P(X = X_i)\, P(Y = Y_j) \qquad [10.2]$$

Dicho con palabras, el suceso «fila = i» es independiente del suceso «columna = j» cuando la probabilidad del suceso «casilla = ij» es igual al producto de las probabilidades del suceso «fila = i» y del suceso «columna = j».

Llamando π_{i+} a la probabilidad de que una observación cualquiera pertenezca a la categoría i de la variable X, π_{+j} a la probabilidad de que una observación cualquiera pertenezca a la categoría j de la variable Y, y π_{ij} a la probabilidad de que una observación cualquiera pertenezca a una de las $I \times J$ casillas, la ecuación [10.2] puede formularse como:

$$\boldsymbol{\pi_{ij} \;=\; \pi_{i+}\, \pi_{+j}} \qquad [10.3]$$

Esto significa que, si se asume que las variables X e Y son independientes, la probabilidad de encontrar una observación cualquiera en una casilla determinada es igual al producto de las

---

3 Aunque *relación* y *asociación* son términos equivalentes en estadística, cuando se trabaja con variables categóricas es habitual utilizar el término *asociación* en lugar del término *relación*, el cual suele reservarse, aunque no necesariamente, para cuando se trabaja con variables cuantitativas. Por otro lado, el concepto de *asociación* entre variables categóricas admite más de un significado teórico; aquí estamos prestando atención únicamente a su significado más extendido. Éste es el significado que nos permite afirmar que, con variables categóricas, *comparar y relacionar son la misma cosa*: decimos que *hay relación* entre las variables cuando *hay diferencias* entre las distribuciones condicionales; decimos que *no hay relación* entre las variables cuando *no hay diferencias* entre las distribuciones condicionales.

4 Por ejemplo, al lanzar al aire dos monedas independientemente, la probabilidad conjunta del suceso *cara en las dos monedas* es igual al producto de las probabilidades individuales de los sucesos *cara en la primera moneda* y *cara en la segunda moneda* ($0{,}5 \times 0{,}5 = 0{,}25$). Ver, en el apartado *Probabilidad: Regla de la multiplicación* del Capítulo 2, la definición de sucesos independientes.

probabilidades marginales de esa casilla. Por tanto, los valores π_{ij} de [10.3] indican lo que cabe esperar encontrar en cada casilla cuando las variables *X* e *Y* son independientes. Si las variables *X* e *Y* no son independientes, la igualdad [10.3] no se sostiene.

La prueba X^2 de Pearson sobre independencia

Ya sabemos que los totales de las columnas de la Tabla 10.3 reflejan el comportamiento de la variable *tabaquismo*: 30,0% de fumadores, 6,5% de exfumadores, 63,5% de no fumadores. Y de lo dicho en el apartado anterior se deduce que, si la variable *tabaquismo* fuera independiente de la variable *sexo*, tanto en el grupo de hombres como en el de mujeres debería reproducirse esa misma pauta (30,0%, 6,5%, 63,5%). Pero está claro que no es esto lo que ocurre: en el grupo de hombres, el porcentaje de fumadores baja hasta el 19,1% y el de no fumadores sube hasta el 73,4%; mientras que en el grupo de mujeres, el porcentaje de fumadoras sube hasta el 39,6% y el de no fumadoras baja hasta el 54,7%.

Ya hemos hecho una interpretación descriptiva de estas diferencias entre las dos distribuciones condicionales de la variable *tabaquismo*. Ahora se trata de ir más allá intentando dilucidar si esas diferencias son trasladables a la población. Es decir, ahora nos preguntamos si las diferencias observadas entre esas dos distribuciones condicionales son lo bastante pequeñas como para pensar que pueden ser atribuidas simplemente a las fluctuaciones propias del azar muestral o, por el contrario, son lo bastante grandes como para reflejar verdaderas diferencias en la población.

Responder a esta pregunta requiere poner a prueba la hipótesis de independencia entre ambas variables: el rechazo de esta hipótesis permitiría concluir que existe relación y que, consecuentemente, las distribuciones condicionales difieren de las marginales. Ahora bien, recordemos que para contrastar una hipótesis es necesario poder establecer el grado de compatibilidad existente entre esa hipótesis y los datos; y esto exige, por un lado, conocer cómo pronostica esa hipótesis que deben comportarse los datos y, por otro, cómo se comportan de hecho. Esto, referido a la hipótesis de independencia en una tabla de contingencias, significa conocer, por un lado, qué valores *cabe esperar que tomen* las frecuencias de la tabla cuando las variables son independientes y, por otro, qué valores *toman de hecho*.

Los valores que toman de hecho se conocen al recoger los datos y construir la tabla de contingencias: son las frecuencias conjuntas que hemos representado mediante n_{ij} (ver Tabla 10.3) y que, a partir de ahora, llamaremos **frecuencias observadas**.

Y los valores que cabe esperar que tomen bajo la condición de independencia se derivan de la ecuación 10.3. Los valores π_{ij} tal como están definidos en [10.3] representan la probabilidad teórica asociada a cada casilla de una tabla de contingencias cuando se asume que las variables que definen la tabla son independientes. Consecuentemente, si se asume independencia, lo que cabe esperar que ocurra al repartir aleatoriamente *n* casos en las $I \times J$ casillas de una tabla de contingencias, viene dado por

$$m_{ij} = n\pi_{ij} = n\pi_{i+}\pi_{+j} \qquad [10.4]$$

Por tanto, los valores m_{ij}, tal como están definidos en [10.4], son los pronósticos que se derivan de la hipótesis de independencia, es decir, las frecuencias que cabe esperar encontrar en cada casilla de una tabla de contingencias cuando las variables que definen la tabla son independientes entre sí. A estos pronósticos los llamaremos **frecuencias esperadas**.

Los valores π_{i+} y π_{+j} son parámetros y, por tanto, generalmente serán valores desconocidos que habrá que estimar a partir de los datos muestrales. Esto puede hacerse sustituyendo las probabilidades teóricas por sus correspondientes empíricas (ver Capítulo 7):

$$P_{i+} = \hat{\pi}_{i+} = \frac{n_{i+}}{n} \qquad \text{y} \qquad \hat{\pi}_{+j} = P_{+j} = \frac{n_{+j}}{n} \qquad [10.5]$$

Y teniendo en cuenta [10.3] y [10.4] se obtiene

$$\hat{m}_{ij} = n P_{i+} P_{+j} = n \frac{n_{i+}}{n} \frac{n_{+j}}{n} = \frac{n_{i+} n_{+j}}{n} \qquad [10.6]$$

Es decir, si se asume independencia ente X e Y, la frecuencia esperada de cada casilla puede estimarse multiplicando sus correspondientes frecuencias marginales y dividiendo ese producto entre el número total de casos.

Para contrastar la hipótesis de independencia sólo falta comparar los pronósticos que se derivan de esa hipótesis (las frecuencias esperadas) con los datos realmente obtenidos (las frecuencias observadas). Esto puede hacerse de distintas maneras, pero lo habitual es utilizar un estadístico, ideado por Pearson (1900, 1911), que adopta la siguiente forma:

$$X^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(n_{ij} - \hat{m}_{ij})^2}{\hat{m}_{ij}} \qquad [10.7]$$

Este estadístico suele encontrarse en la literatura estadística con el nombre ***ji*-cuadrado**[5]. Su valor oscila entre cero e infinito. Cuando los datos se comportan exactamente tal como pronostica la hipótesis de independencia, las diferencias entre las frecuencias observadas y sus correspondientes esperadas valen, todas ellas, cero; y el estadístico de Pearson también vale cero. Su valor va aumentando, alejándose de cero, tanto más cuanto más difieren los datos de los pronósticos basados en la hipótesis de independencia, es decir, cuanto mayores son las diferencias entre las frecuencias observadas y las esperadas.

La cuestión clave está en determinar cuándo el valor del estadístico de Pearson se aleja de cero lo bastante como para decidir que la hipótesis de independencia es falsa. Lo cual tiene fácil solución porque conocemos la distribución muestral del estadístico de Pearson:

$$X^2 \sim \chi^2_{(I-1)(J-1)} \qquad [10.8]$$

Es decir, el estadístico X^2 de Pearson se distribuye, aproximadamente, según el modelo teórico de probabilidad χ^2 (*ji*-cuadrado; ver Apéndice 5) con $(I-1)(J-1)$ grados de libertad[6].

Llegados a este punto, sabemos qué hipótesis queremos contrastar (la hipótesis de independencia entre dos variables categóricas), sabemos qué pronósticos se derivan de ella (las frecuencias esperadas) y con qué tipo de datos compararlos (las frecuencias observadas), y

---

[5] Este estadístico también puede encontrarse en la literatura estadística en español con el nombre *chi-cuadrado* (anglicismo innecesario e incorrecto que, no obstante, utilizaremos en algún momento por ser el que aparece en el SPSS).

[6] Puesto que todas las casillas de la tabla contribuyen al resultado del estadístico de Pearson, parecería lógico pensar que su distribución debería tener tantos grados de libertad como casillas tiene la tabla. Pero en una tabla de contingencias las frecuencias esperadas dependen del valor de las frecuencias marginales. Por tanto, no todas las frecuencias esperadas aportan información nueva; algunas son redundantes. En una tabla de contingencias, los grados de libertad indican el número de casillas cuyas frecuencias esperadas pueden adoptar libremente cualquiera de sus posibles valores tras fijar el valor de las frecuencias marginales.

tenemos un estadístico que, además de permitir efectuar esas comparaciones, tiene distribución muestral conocida. Es decir, tenemos todo lo necesario para poder diseñar un contraste de la hipótesis de independencia ente dos variables categóricas. El Cuadro 10.1 ofrece un resumen de este contraste.

Cuadro 10.1. Resumen del contraste de hipótesis sobre independencia entre dos variables categóricas (prueba X^2 de Pearson sobre independencia)

1. *Hipótesis*:
   H_0: X e Y son variables independientes (Es decir, $\pi_{ij} = \pi_{i+}\,\pi_{+j}$, para todo ij).
   H_1: X e Y no son variables independientes (Es decir, $\pi_{ij} \neq \pi_{i+}\,\pi_{+j}$, para algún ij).
2. *Supuestos*: una muestra aleatoria de n observaciones es clasificada en las $I \times J$ combinaciones (casillas) resultantes de combinar dos variables categóricas; la probabilidad de que una observación cualquiera pertenezca a cada una de las casillas se mantiene constante durante todo el proceso de clasificación; no más del 20% de las frecuencias esperadas son menores que 5 (ver, en el Apéndice 9, el apartado *Supuestos del estadístico X^2 de Pearson*).
3. *Estadístico del contraste* (ver 10.7]: $X^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(n_{ij} - \hat{m}_{ij})^2}{\hat{m}_{ij}}$
4. *Distribución muestral*: X^2 se aproxima a χ^2 con $(I-1)(J-1)$ grados de libertad conforme n va aumentando.
5. *Zona crítica*: $x^2 \geq \chi^2_{(I-1)(J-1);\,1-\alpha}$ (ver Tabla D del Apéndice final).
6. *Regla de decisión*: se rechaza H_0 si el estadístico X^2 cae en la zona crítica; en caso contrario, se mantiene. Si se rechaza H_0 se concluye que las variables X e Y están relacionadas.
7. *Nivel crítico* (valor p): $p = P(X^2 \geq X_h^2)$, donde X_h^2 se refiere al valor muestral concreto que toma X^2.

Ejemplo. La prueba X^2 de Pearson sobre independencia

Veamos cómo contrastar la hipótesis de independencia con los mismos datos utilizados para explicar el concepto de asociación en tablas de contingencias, es decir, con los datos de la Tabla 10.2. Recordemos que esa tabla contiene el resultado de clasificar una muestra de 200 personas según dos criterios: *sexo* (hombres, mujeres) y *tabaquismo* (fumadores, exfumadores y no fumadores). La pregunta que nos hacemos es la siguiente: ¿podemos afirmar que las variables *sexo* y *tabaquismo* están relacionadas? ($\alpha = 0{,}05$).

1. *Hipótesis*:
   H_0: las variables *sexo* y *tabaquismo* son independientes.
   H_1: las variables *sexo* y *tabaquismo* no son independientes (están relacionadas).

2. *Supuestos*: una muestra aleatoria de 200 sujetos se ha clasificado en las 6 combinaciones (casillas) exclusivas y exhaustivas resultantes de combinar las variables *sexo* y *tabaquismo*. La probabilidad de que una observación pertenezca a cada una de las casillas se mantiene constante durante todo el proceso de clasificación.

3. *Estadístico del contraste*. Para obtener el estadístico de Pearson es necesario calcular previamente las frecuencias esperadas que se derivan de la hipótesis de independencia. Para ello utilizamos la estrategia propuesta en la ecuación [10.6]. La Tabla 10.6 ofrece estas frecuencias esperadas, entre paréntesis, junto a las observadas. Por ejemplo, las frecuencias esperadas de las casillas de la primera columna de la tabla (hombres-fumadores y mujeres-fumadoras) se ha obtenido de la siguiente manera:

$$\hat{m}_{11} = \frac{n_{1+}n_{+1}}{n} = \frac{94\,(60)}{200} = 28{,}2$$

$$\hat{m}_{21} = \frac{n_{2+}n_{+1}}{n} = \frac{106\,(60)}{200} = 31{,}8$$

Tabla 10.6. Tabla de contingencias de *sexo* por *tabaquismo*: frecuencias observadas (esperadas)

	Tabaquismo			
Sexo	Fumadores	Exfumadores	No fumadores	*Total*
Hombres	18 (28,2)	7 (6,1)	69 (59,7)	94
Mujeres	42 (31,8)	6 (6,9)	58 (67,3)	106
Total	60	13	127	200

No hay frecuencias esperadas menores que 5.

Aplicando el estadístico de Pearson a estas frecuencias se obtiene

$$X^2 = \sum_i \sum_j \frac{(n_{ij} - \hat{m}_{ij})^2}{\hat{m}_{ij}} = \frac{(18-28{,}2)^2}{28{,}2} + \frac{(7-6{,}1)^2}{6{,}1} + \cdots + \frac{(58-67{,}3)^2}{67{,}3} = 9{,}95$$

4. *Distribución muestral*: χ^2 con $(I-1)(J-1) = (2-1)(3-1) = 2$ grados de libertad: χ^2_2.

5. *Zona crítica*: $X^2 \geq \chi^2_{2;\,0,95} = 7{,}38$.

6. *Decisión*: puesto que el valor de estadístico del contraste (9,95) es mayor que el punto crítico (5,99), se rechaza H_0. Se puede concluir, por tanto, que las variables *sexo* y *tabaquismo* no son independientes; o, lo que es lo mismo, que están relacionadas.

7. *Nivel crítico* (valor *p*): en la distribución *ji*-cuadrado con 2 grados de libertad (Tabla D del Apéndice final), $p = P(X^2 \geq 9{,}95) < 0{,}01$.

   Puesto que la tabla sólo ofrece algunos cuantiles, normalmente no será posible calcular a partir de ella el nivel crítico exacto. No obstante, dado que el objetivo de calcular el nivel crítico es el de poder tomar una decisión sobre H_0, normalmente bastará con saber si su valor es menor o mayor que α. Si se desea obtener el nivel crítico exacto, puede utilizarse, en SPSS, la función CDF.CHISQ (ver Apéndice 5).

Medidas de asociación

Ya sabemos cómo describir dos variables categóricas mediante tablas de contingencias y gráficos de barras agrupadas. También sabemos cómo contrastar la hipótesis de independencia con la prueba X^2 de Pearson para decidir si las variables están o no relacionadas. Nos falta saber qué hacer cuando se rechaza la hipótesis de independencia; es decir, nos falta saber cómo *cuantificar* la relación encontrada (se explica en este apartado) y qué hacer para *interpretarla* (se explica en el siguiente apartado).

Aunque, según hemos visto, el estadístico X^2 de Pearson permite contrastar la hipótesis de independencia, no contiene información sobre la *fuerza* o *intensidad* de la asociación. Esto se debe a que su valor no depende únicamente del grado de asociación real existente entre las variables sino, también, del tamaño muestral. Con tamaños muestrales muy grandes, diferencias relativamente pequeñas entre las frecuencias observadas y las esperadas pueden dar lugar a valores X^2 muy altos sin que esto implique necesariamente una fuerte asociación. Para cuantificar correctamente la fuerza o intensidad de la asociación es necesario utilizar estadísticos capaces de eliminar el efecto del tamaño muestral. Estos estadísticos se conocen como **medidas de asociación**.

Existen diversas medidas que difieren no sólo en la forma de definir lo que es asociación, sino en la forma en que cada una se ve afectada por factores tales como las distribuciones marginales o la naturaleza de las variables que definen la tabla. Las medidas de asociación incluidas en este apartado se basan en el estadístico X^2 de Pearson (estas medidas son *nominales* en el sentido de que únicamente aprovechan información nominal; con variables nominales no tiene sentido hablar de la dirección o naturaleza de una asociación; en el segundo volumen se estudian otras medidas de asociación). Todas ellas intentan cuantificar el grado de asociación aplicando algún tipo de corrección al valor del estadístico X^2 para hacerle tomar un valor comprendido entre 0 y 1. La primera de estas medidas es el **coeficiente de contingencia *C*** (Pearson, 1913),

$$C = \sqrt{X^2/(X^2+n)} \qquad [10.9]$$

Toma valores comprendidos entre 0 y un máximo que siempre es menor que 1 (puesto que n siempre es mayor que 0, C nunca puede llegar a 1). Este máximo depende del número de filas y de columnas de la tabla. Si el número de filas y de columnas es el mismo (k), el valor máximo de C se obtiene mediante $C_{\text{máx.}} = \sqrt{(k-1)/k}$. En condiciones de independencia total (frecuencias condicionales idénticas) el coeficiente C vale 0; en condiciones de asociación perfecta (máxima diferencia entre distribuciones condicionales) el coeficiente toma su valor máximo. Con los datos del ejemplo anterior (donde X^2 = 9,95 y n = 200), se obtiene

$$C = \sqrt{9{,}95/(9{,}95+200)} = 0{,}22$$

Otra medida de asociación basada en X^2 y muy parecida al coeficiente de contingencia C es el **coeficiente *V* de Cramèr** (1946):

$$V_{\text{Cramèr}} = \sqrt{X^2/[n\,(k-1)]} \qquad [10.10]$$

Ahora k se refiere al valor más pequeño del número de filas y de columnas. Al igual que C, el coeficiente V toma el valor 0 en condiciones de independencia total; pero, a diferencia de

lo que ocurre con C, en condiciones de asociación perfecta el coeficiente V toma el valor 1 independientemente del número de filas y de columnas de la tabla. Con los datos del ejemplo anterior, se obtiene

$$V_{\text{Cramèr}} = \sqrt{9{,}95/[200\,(2-1)]} = 0{,}22$$

En tablas de contingencia 2×2, el coeficiente $V_{\text{Cramér}}$ se reduce a

$$\varphi = \sqrt{X^2/n} \qquad [10.11]$$

Esta expresión se conoce como **coeficiente phi** (φ), el cual no es otra cosa que el coeficiente de correlación de Pearson (ver Capítulo 12) aplicado a dos variables dicotómicas codificadas con "unos" y "ceros".

Estas medidas de asociación son muy útiles para comparar la relación entre dos variables cuando ésta se calcula en grupos o momentos distintos. Pero no es fácil interpretarlas tomando como único referente la tabla en la que se calculan. El contexto es importante. Podría pensarse que un coeficiente de 0,40 está indicando que el grado de relación es bajo-medio. Pero esto es algo que habría que referir al contexto en el que se observa esa relación. Por ejemplo, si cada vez que se correlaciona una variable (cualquier variable) con otras se obtienen cuantificaciones que no pasan de 0,20, encontrar una relación de 0,40 puede representar un hallazgo importante.

Residuos tipificados

Poder concluir que dos variables categóricas están asociadas es, sin duda, interesante. Pero lo realmente interesante es poder aclarar el significado de la asociación encontrada. Para ello, basta con tener presente que si se rechaza la hipótesis de independencia es porque, al menos en una casilla, se produce una diferencia importante entre lo observado y lo esperado.

Por tanto, una vez rechazada la hipótesis de independencia, la pauta de asociación concreta presente en una tabla de contingencias pueden estudiarse realizando una valoración casilla a casilla de las diferencias existentes entre las frecuencias esperadas y las observadas. A estas diferencias se les llama *residuos*:

$$R_{ij} = n_{ij} - \hat{m}_{ij} \qquad [10.12]$$

Los residuos pueden delatar diferencias mayores en unas celdas que en otras y la constatación de este hecho puede arrojar luz sobre la pauta de asociación presente en la tabla. Una forma sencilla de evaluar estos residuos consiste en tipificarlos:

$$Z_{R_{ij}} = \frac{R_{ij}}{\sqrt{\hat{m}_{ij}}} \qquad [10.13]$$

Los residuos tipificados elevados al cuadrado poseen la interesante propiedad de ser componentes del estadístico de Pearson: $\sum_i \sum_j Z^2_{R_{ij}} = X^2$. Bajo la hipótesis de independencia se distribuyen normalmente. Pero, aunque su media vale 0, su varianza es $(I-1)(J-1)/(I \times J)$, lo cual representa un pequeño inconveniente pues, dado que $(I-1)(J-1)$ es siempre menor que

$I \times J$, la varianza de los residuos tipificados siempre es menor que 1 y, en consecuencia, su variabilidad no se corresponde exactamente con la de las variables distribuidas $N(0, 1)$.

No obstante, Haberman (1973) ha definido otro tipo de residuos tipificados, llamados *ajustados* o *corregidos*, que, a diferencia de los tipificados, sí se distribuyen $N(0, 1)$. Se calculan dividiendo el residuo de cada casilla por su *error típico*:

$$Z_{R_{ij\,(\text{corregido})}} = R_{ij} / \sqrt{\hat{m}_{ij}\,(1 - n_{i+}/n)\,(1 - n_{+j}/n)} \qquad [10.14]$$

En el ejemplo de la Tabla 10.6, el residuo tipificado corregido correspondiente a la primera casilla de la tabla se obtiene de la siguiente manera:

$$Z_{R_{12\,(\text{corregido})}} = (18 - 28{,}2) / \sqrt{28{,}2\,(1 - 94/200)\,(1 - 60/200)} = -3{,}15$$

La gran utilidad de los residuos tipificados corregidos radica precisamente en que su distribución es normal con media cero y desviación típica uno: $N(0, 1)$. Una variable de estas características es fácilmente interpretable: utilizando un nivel de confianza de, por ejemplo, 0,95, puede afirmarse que los residuos mayores que 1,96 (puntuación típica correspondiente al cuantil 97,5 en una distribución normal) delatan casillas con más casos de los que cabría esperar por azar si las dos variables estudiadas fueran realmente independientes; mientras que los residuos menores que −1,96 (puntuación típica correspondiente al cuantil 2,5 en una distribución normal) delatan casillas con menos casos de los que cabría esperar si las dos variables estudiadas fueran realmente independientes. Por tanto, una vez rechazada la hipótesis de independencia mediante el estadístico de Pearson, los residuos tipificados corregidos constituyen una herramienta muy útil para poder interpretar con precisión el significado de la asociación detectada, permitiendo valorar hacia dónde y desde dónde se producen desplazamientos significativos de casos.

El valor −3,15 obtenido para el residuo tipificado corregido de la primera casilla de la Tabla 10.2 indica que, en esa casilla, la frecuencia observada es significativamente menor que la esperada (pues $-3{,}15 < -1{,}96$); lo cual permite afirmar que existen menos hombres fumadores de los que pronostica la hipótesis de independencia: donde cabía esperar un 30,0% de hombres fumadores, se ha encontrado un 19,1%. El valor −3,1 indica que esos dos porcentajes difieren significativamente, es decir, que difieren más de lo que cabría esperar sólo por azar si realmente el porcentaje de hombres fumadores fuera del 30,0%.

Tablas de contingencias y gráficos de barras con SPSS

El procedimiento **Tablas de contingencias** permite obtener, además de la prueba X^2 de Pearson sobre independencia, las tablas de contingencias y los gráficos de barras agrupadas estudiados en este capítulo.

Las tablas de contingencias que genera el SPSS por defecto contienen únicamente las frecuencias observadas. No obstante, el procedimiento incluye opciones que permiten solicitar las frecuencias esperadas y los tres tipos de frecuencias porcentuales que hemos estudiado: las referidas a las filas, a las columnas y al total (no debe olvidarse que son las frecuencias porcentuales de las filas y de las columnas las que realmente informan acerca de la relación entre las variables).

El SPSS construye los gráficos de barras agrupadas colocando en el eje horizontal las categorías de la variable *fila* y, sobre ellas, anidadas, las de la variable *columna*. Cada barra del gráfico corresponde, por tanto, a una casilla de la tabla; y la altura de las barras representa el tamaño de las frecuencias observadas. El procedimiento **Tablas de contingencias** no permite obtener los gráficos de barras agrupadas correspondientes a los porcentajes de las filas y de las columnas (que son, no lo olvidemos, los que realmente reflejan el significado de una relación); para obtener estos gráficos es necesario recurrir a los gráficos de barras disponibles en el menú **Gráficos**.

Puede seleccionarse más de una variable *fila* y más de una variable *columna*; en ese caso, cada variable *fila* se cruza con cada variable *columna* para formar una tabla distinta. Por ejemplo, con 2 variables *fila* y 3 variables *columna*, se obtienen $2 \times 3 = 6$ tablas distintas.

Para poder trabajar con los datos de la Tabla 10.2, vamos a comenzar reproduciéndolos en el *Editor de datos* tal como muestra la Figura 10.3 (los datos se encuentran en el archivo *Tabla 10.2 sexo por tabaco*, en la página web del manual). Hemos creado las variables *sexo* (con etiqueta *Sexo*), *tabaco* (con etiqueta *Tabaquismo*) y *ncasos* (con las frecuencias asociadas a cada casilla). Para que los 6 casos del archivo representado en la Figura 10.3 se conviertan en los 200 de la Tabla 10.2 es necesario ponderar el archivo con la variable *ncasos* (esto se hace con la opción **Ponderar casos** del menú **Datos**).

Tabla 10.2 (bis). Tabla de contingencias de *sexo* por *tabaquismo*

	Tabaquismo			
Sexo	Fumadores	Exfumadores	No fumadores	*Total*
Hombres	18	7	69	94
Mujeres	42	6	58	106
Total	60	13	127	200

Figura 10.3. Datos de la Tabla 10.2 reproducidos en el *editor de datos*

	sexo	tabaco	ncasos
1	1	1	18
2	1	2	7
3	1	3	69
4	2	1	42
5	2	2	6
6	2	3	58

Para obtener la tabla de contingencias y el diagrama de barras agrupadas correspondientes a los datos de la Tabla 10.2 (archivo de la Figura 10.3):

➤ Seleccionar opción **Estadísticos descriptivos > Tablas de contingencias** del menú **Analizar**, trasladar la variable *sexo* a la lista **Filas** y la variable *tabaco* a la lista **Columnas** y marcar la opción **Mostrar los gráficos de barras agrupadas**.

➤ Pulsar el botón **Casillas** para acceder al subcuadro de diálogo **Tablas de contingencias: Mostrar en las casillas** y marcar las opciones **Fila**, **Columna** y **Total** del apartado **Porcentajes**. Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.

Aceptando estas selecciones el *Visor de resultados* ofrece la tabla de contingencias de *sexo* por *tabaquismo* y el correspondiente gráfico de barras agrupadas. La Tabla de contingencias está reproducida en la Tabla 10.7. Además de las frecuencias observadas de la Tabla 10.2 (*recuento*), la tabla incluye los porcentajes de fila de la Tabla 10.4 (*% de sexo*) y los porcentajes de columna de la Tabla 10.5 (*% de tabaquismo*). El significado de estas frecuencias porcentuales ya se ha explicado más arriba en el apartado *Tipos de frecuencias*.

El gráfico de barras agrupadas que ofrece el procedimiento ya se ha presentado y discutido en la Figura 10.1 (derecha). No se debe olvidar que este gráfico, aunque tiene utilidad descriptiva (pues refleja la frecuencia con la que se da cada combinación de categorías), no contiene información útil para interpretar la relación entre las variables.

Tabla 10.7. Tabla de contingencias de *sexo* por *tabaquismo*

			Tabaquismo			
			Fumadores	Exfumadores	No fumadores	Total
Sexo	Hombres	Recuento	18	7	69	94
		% de Sexo	19,1%	7,4%	73,4%	100,0%
		% de Tabaquismo	30,0%	53,8%	54,3%	47,0%
		% del total	9,0%	3,5%	34,5%	47,0%
	Mujeres	Recuento	42	6	58	106
		% de Sexo	39,6%	5,7%	54,7%	100,0%
		% de Tabaquismo	70,0%	46,2%	45,7%	53,0%
		% del total	21,0%	3,0%	29,0%	53,0%
Total		Recuento	60	13	127	200
		% de Sexo	30,0%	6,5%	63,5%	100,0%
		% de Tabaquismo	100,0%	100,0%	100,0%	100,0%
		% del total	30,0%	6,5%	63,5%	100,0%

La prueba X^2 de Pearson sobre independencia con SPSS

Este apartado muestra cómo contrastar la hipótesis de independencia con la prueba X^2 de Pearson y cómo interpretar una relación significativa a partir de los residuos tipificados corregidos. Seguimos trabajando con los datos de la Tabla 10.2 (reproducidos en el *Editor de datos* tal como muestra la Figura 10.3). Para obtener el estadístico de Pearson:

➤ Seleccionar opción **Estadísticos descriptivos** > **Tablas de contingencias** del menú **Analizar** y trasladar la variable *sexo* a la lista **Filas** y la variable *tabaco* a la lista **Columnas**.

➤ Pulsar el botón **Estadísticos** para acceder al subcuadro de diálogo **Tablas de contingencias: Estadísticos** y marcar la opción **Chi-cuadrado**. Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.

➤ Pulsar el botón **Casillas** y marcar la opción **Tipificados corregidos** del apartado **Residuos**. Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.

Aceptando estas selecciones, el *Visor* ofrece los resultados que muestran las Tablas 10.8 y 10.9. Comencemos con la 10.9. El estadístico de Pearson (*chi-cuadrado*) aparece en la primera fila de la tabla. Su valor es 9,945 y tiene asociados 2 grados de libertad (*gl*). En la distribu-

ción χ^2 con 2 grados de libertad, el valor 9,945 tiene asociada una probabilidad (*sig. asintótica bilateral*) de 0,007 (ésta es la probabilidad que queda por encima o a la derecha de 9,945). Dado que esta probabilidad (denominada *valor p*, *nivel crítico* o *nivel de significación observado*) es menor que 0,05, se puede rechazar la hipótesis de independencia y concluir que las variables *sexo* y *tabaquismo* están relacionadas. La tabla incluye otros estadísticos a los que, de momento, no prestaremos atención.

La Tabla 10.8 muestra, además de la frecuencia observada de cada casilla, el correspondiente residuo tipificado corregido (*residuo corregido*). Éstos son los valores que ayudan a interpretar el significado de la relación encontrada (recordemos que estos valores hay que compararlos con 1,96 y −1,96, es decir, con los cuantiles 2,5 y 97,5 de la distribución normal tipificada). Los resultados obtenidos permiten afirmar que, entre los hombres hay menos fumadores (−3,2 < −1,96) y más no fumadores (2,7 > 1,96) de los que pronostica la hipótesis de independencia, mientras que entre las mujeres hay más fumadoras (3,2 > 1,96) y menos no fumadoras (−2,7) de las que pronostica la hipótesis de independencia. Por tanto, el porcentaje de fumadores es menor y el de fumadoras mayor que el esperado (30,0%); y el porcentaje de no fumadores es mayor y el de no fumadoras menor que el esperado (65,5%).

Tabla 10.8. Tabla de contingencias de *sexo* por *tabaquismo* con los residuos tipificados corregidos

			Tabaquismo			
			Fumadores	Exfumadores	No fumadores	Total
Sexo	Hombres	Recuento	18	7	69	94
		Residuos corregidos	-3,2	,5	2,7	
	Mujeres	Recuento	42	6	58	106
		Residuos corregidos	3,2	-,5	-2,7	
Total		Recuento	60	13	127	200

Tabla 10.9. Estadísticos

	Valor	gl	Sig. asintótica (bilateral)
Chi-cuadrado de Pearson	9,945[a]	2	,007
Razón de verosimilitudes	10,184	2	,006
Asociación lineal por lineal	9,240	1	,002
N de casos válidos	200		

a. 0 casillas (0%) tienen una frecuencia esperada inferior a 5. La frecuencia esperada más pequeña es 6,11.

Apéndice 10

Tablas de contingencias con variables de respuesta múltiple

Además de las tablas de frecuencias ya estudiadas en el Apéndice 3, el SPSS permite obtener tablas de contingencias combinando conjuntos de respuestas múltiples con la opción **Respuestas múltiples > Tablas de contingencias** del menú **Analizar**. Por supuesto, antes de poder obtener tablas de contingencias es necesario definir los correspondientes conjuntos de respuestas múltiples tal como se ha explicado en el

Apéndice 3. En el cuadro de diálogo **Tablas de contingencias de repuestas múltiples**, la lista de variables del archivo de datos ofrece un listado de todas las variables con formato numérico; y la lista **Conjuntos de respuestas múltiples** ofrece un listado de los conjuntos previamente definidos.

El procedimiento permite obtener tablas de contingencias de dos dimensiones (combinando dos conjuntos o variables) y de tres dimensiones (combinando tres conjuntos o variables). Para obtener una tabla de contingencias de *dos dimensiones*: trasladar un conjunto (o variable) a la lista **Filas** y un segundo conjunto (o variable) a la lista **Columnas**. Es posible combinar tanto un conjunto con otro como un conjunto con una variable individual (también es posible combinar dos variables individuales, pero para esto es preferible utilizar el procedimiento **Tablas de contingencias** ya explicado en este mismo capítulo). Para obtener una tabla de contingencias de *tres dimensiones* hay que trasladar un tercer conjunto (o variable) a la lista **Capas**.

Cuando se traslada una variable individual (no un conjunto) a cualquiera de las dimensiones del cuadro de diálogo (filas, columnas, capas), el SPSS coloca detrás de la variable dos signos de interrogación entre paréntesis. Esto significa que el procedimiento está esperando que se defina, para esa variable, el rango de valores que se desea incluir en la tabla de contingencias. Para definir el rango de valores de una variable: seleccionar la variable cuyo rango se desea definir y pulsar el botón **Definir rangos** e introducir, en los cuadros de texto **Mínimo** y **Máximo**, los códigos más pequeño y más grande del rango de códigos que identifican a las categorías que se desea tabular.

Si no se indica otra cosa, el procedimiento ofrece las frecuencias de respuesta conjunta únicamente en valor absoluto, no en valor porcentual. Y las frecuencias marginales (en valor absoluto y porcentual) las calcula tomando como referencia el número de casos (no en el de respuestas). No obstante, el botón **Opciones** conduce a un subcuadro de diálogo que permite decidir si se desea o no obtener frecuencias porcentuales y si éstas deben basarse en el número de casos válidos del archivo o en el número de respuestas:

1. Las opciones del recuadro **Porcentajes de casilla** permiten decidir qué tipo de porcentajes se desea incluir en las casillas de la tabla de contingencias: la opción **Fila** ofrece el porcentaje que la frecuencia de respuesta de cada casilla representa sobre el total de casos o respuestas de su fila; la opción **Columna** ofrece el porcentaje que la frecuencia de respuesta de cada casilla representa sobre el total de casos o respuestas de su columna; la opción **Total** ofrece el porcentaje que la frecuencia de respuesta de cada casilla representa sobre el total de casos o respuestas de la tabla.

   Al combinar dos conjuntos de respuestas múltiples, el SPSS cruza, por defecto, cada variable del primer conjunto con cada variable del segundo conjunto y suma las respuestas. Cuando se están combinando conjuntos de *categorías múltiples*, la opción **Emparejar las variables entre los conjuntos de respuesta** hace que la primera variable del primer conjunto se cruce con la primera variable del segundo conjunto, la segunda variable del primer conjunto con la segunda variable del segundo conjunto, etc.; de esta manera, el número de respuestas computadas es menor que el número total de respuestas. Al marcar esta opción, los porcentajes de respuesta únicamente es posible obtenerlos tomando como referencia el número de respuestas (no el número de casos); de hecho, al marcar esta opción se desactiva la opción **Casos** del recuadro **Porcentajes basados en**.

2. Las opciones del recuadro **Porcentajes basados en** permiten elegir el referente sobre el cual se calculan los porcentajes: si se elige la opción **Casos**, el referente es el número de casos válidos; si se elige la opción **Respuestas**, el referente es el número de respuestas. Debe tenerse en cuenta que, al trabajar con variables de respuesta múltiple, el número de respuestas es mayor que el número de casos. En los conjuntos de dicotomías múltiples, el número de respuestas se obtiene a partir del número de "unos" (valor contado) de cada dicotomía. En los conjuntos de categorías múltiples, el número de respuestas se obtiene a partir del número de respuestas con valor comprendido en el rango establecido al definir el conjunto.

3. Las opciones del recuadro **Valores perdidos** permiten controlar el tipo de tratamiento que se desea dar a los valores perdidos. Es posible excluir del análisis los casos con valor perdido en cualquiera de los conjuntos seleccionados o excluir únicamente los casos con valor perdido en cada conjunto

analizado. Un caso se considera un valor perdido cuando no puntúa (valor contado) en ninguna de las variables del conjunto definido como dicotomías múltiples o cuando no toma un valor comprendido dentro del rango de valores definido para el conjunto de categorías múltiples.

Cómo obtener tablas de contingencias

Veamos cómo obtener tablas de contingencias con conjuntos de respuestas múltiples. Para ello, vamos a utilizar la variable *sexo* (variable individual: 1 = «hombres», 2 = «mujeres») y el conjunto de dicotomías múltiples *trans_d* definido en el Apéndice 3 (ver Figura 3.6). Para cruzar estas dos variables en una tabla de contingencias:

- ➤ Seleccionar la opción **Respuestas múltiples > Tablas de contingencias** del menú **Analizar** y trasladar la variable *sexo* a la lista **Filas**.
- ➤ Manteniendo seleccionada la variable *sexo* dentro de la lista **Filas**, pulsar el botón **Definir rangos**, introducir los códigos 1 y 2 en los cuadros de texto **Mínimo** y **Máximo**, y pulsar el botón **Continuar** para volver al cuadro de diálogo principal.
- ➤ Seleccionar el conjunto *$trans_d* (transporte público, dicotomías) y trasladarlo a la lista **Columnas**.
- ➤ Pulsar el botón **Opciones** y marcar las opciones **Fila**, **Columna** y **Total** del recuadro **Porcentajes de casilla**. Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.

Aceptando estas selecciones, el *Visor* ofrece los resultados que muestra la Tabla 10.10. El contenido de las casillas (*recuento*) refleja el *número de respuestas* de cada tipo: 6 *hombres* utilizan el *autobús*, 8 *mujeres* utilizan el *metro*, etc. Los totales de las columnas (que son los totales que corresponden al conjunto *transporte público*) también reflejan el número de respuestas: de los 20 encuestados, 10 manifiestan utilizar el *autobús*, 14 el *metro*, etc.

El resto de valores de la tabla está referido al *número de encuestados* o número de casos válidos (esta circunstancia queda aclarada en una nota a pie de tabla). Así, los totales marginales de las filas (los totales marginales de la variable individual *género*) reflejan el número de hombres y mujeres. Y los porcentajes de las casillas están calculados sobre el número de casos válidos: los 6 *hombres* que utilizan el *metro* representan el 60,0% de los 10 hombres encuestados, el 42,9% de los 14 encuestados que utilizan el metro y el 30,0% de los 20 encuestados de la muestra.

El total de la tabla indica que la muestra está formada por 20 encuestados o casos válidos. Este total coincide con la suma de las frecuencias marginales de las filas (que recogen el número de encuestados), pero no con la suma de las frecuencias marginales de las columnas (que recogen el número de respuestas).

Tabla 10.10. Tabla de contingencias de *género* por *transporte público* (porcentajes basados en casos)

			$trans_d				Total
			Autobús	Metro	Tren	Taxi	
Sexo	Hombres	Recuento	6	6	7	2	10
		% dentro de sexo	60,0%	60,0%	70,0%	20,0%	
		% dentro de $trans_d	60,0%	42,9%	53,8%	40,0%	
		% del total	30,0%	30,0%	35,0%	10,0%	50,0%
	Mujeres	Recuento	4	8	6	3	10
		% dentro de sexo	40,0%	80,0%	60,0%	30,0%	
		% dentro de $trans_d	40,0%	57,1%	46,2%	60,0%	
		% del total	20,0%	40,0%	30,0%	15,0%	50,0%
Total		Recuento	10	14	13	5	20
		% del total	50,0%	70,0%	65,0%	25,0%	100,0%

Los porcentajes y los totales se basan en los encuestados.

Marcando la opción **Respuestas** en el recuadro **Porcentajes basados en** del botón **Opciones**, todos los valores de la tabla (frecuencias y porcentajes) quedan referidos, no al *número de encuestados*, sino al *número de respuestas* (ver Tabla 10.11). Una nota a pie de tabla indica esta circunstancia cambiando de *encuestados* (casos válidos) a *respuestas*. Ahora, los 6 *hombres* que utilizan el *metro* representan un 28,6% de las 21 respuestas dadas por los 10 hombres encuestados, un 42,9% de los 14 encuestados que utilizan el metro y un 14,3% de las 42 respuestas dadas por los 20 encuestados de la muestra.

El total de la tabla indica que los 20 encuestados han dado 42 respuestas. Puesto que ahora todos los valores están basados en el número de respuestas, este total coincide tanto con la suma de los totales de las filas como con la suma de los totales de las columnas.

Tabla 10.11. Tabla de contingencias de *género* por *transporte público* (porcentajes basados en respuestas)

			$trans_d				Total
			Autobús	Metro	Tren	Taxi	
Sexo	Hombres	Recuento	6	6	7	2	21
		% dentro de sexo	28,6%	28,6%	33,3%	9,5%	
		% dentro de $trans_d	60,0%	42,9%	53,8%	40,0%	
		% del total	14,3%	14,3%	16,7%	4,8%	50,0%
	Mujeres	Recuento	4	8	6	3	21
		% dentro de sexo	19,0%	38,1%	28,6%	14,3%	
		% dentro de $trans_d	40,0%	57,1%	46,2%	60,0%	
		% del total	9,5%	19,0%	14,3%	7,1%	50,0%
Total		Recuento	10	14	13	5	42
		% del total	23,8%	33,3%	31,0%	11,9%	100,0%

Los porcentajes y los totales se basan en las respuestas.

Ejercicios

10.1. A continuación se ofrecen algunos datos extraídos de un trabajo realizado en la Comunidad de Madrid en el que, entre otras cosas, se ha estudiado la relación entre la *edad* (agrupada en tres categorías) y la *causa de la muerte* (sólo se consideran algunas causas). Los datos corresponden a una muestra de 2.000 defunciones seleccionadas aleatoriamente entre las registradas en 2007 en personas de 15 a 70 años.

	Grupos de edad			
Causa de la muerte	15-30 años	31-50 años	51-70 años	*Totales*
Accidente	51	93	77	221
Cáncer	6	64	483	553
Enfermedad cardíaca	4	54	321	379
Homicidio	13	23	14	50
Suicidio	14	31	37	82
Otra causa	24	167	524	715
Totales	112	432	1456	2.000

a. ¿Qué porcentaje de personas mueren de cáncer?
b. ¿Qué porcentaje de personas de la muestra total muere por causa de enfermedad cardíaca entre los 51 y los 70 años?
c. En el grupo de edad de 31 a 50 años, ¿qué porcentaje de personas se suicida?
d. Entre las personas que mueren por accidente, ¿qué porcentaje tiene entre 15 y 30 años?
e. ¿Puede afirmarse, con $\alpha = 0{,}05$, que la edad está relacionada con la causa de la muerte?
f. ¿Cuál es la intensidad de la relación?
g. ¿Qué causas de muerte predominan en cada grupo de edad?

10.2. En varios estudios realizados durante los últimos años se ha llegado a conclusiones contradictorias acerca de la relación existente entre el estado civil y la actitud hacia el aborto. Con intención de aportar nueva evidencia empírica sobre esta relación, se ha encuestado a 500 sujetos y, tras clasificarlos según su estado civil y su actitud hacia el aborto, se han obtenido los resultados que aparecen en la siguiente tabla:

	Actitud hacia el aborto	
Estado civil	A favor	En contra
Solteros	130	20
Casados	50	150
Divorciados / separados	40	60
Viudos	20	30

a. Calcular los porcentajes de fila de la tabla.
b. ¿Puede afirmarse, con $\alpha = 0{,}05$, que la actitud hacia el aborto está relacionada con el estado civil?
c. En el caso de que exista relación, interpretarla.

10.3. Se ha diseñado un estudio para analizar la posible relación entre el hábitat y la incidencia de trastorno depresivo en las personas en paro. Se han seleccionado sujetos pertenecientes a medios rurales, semiurbanos y urbanos. De cada medio se ha seleccionado una muestra aleatoria de 100 sujetos en paro, obteniendo en cada grupo el número de depresivos que aparece en la tabla:

	Hábitat		
Trastorno depresivo	Rural	Semiurbano	Urbano
Sí	12	16	32
No	88	84	68

¿Puede afirmarse, con $\alpha = 0{,}01$, que en la población de desempleados existe relación entre el tipo de medio en el que se vive y padecer o no trastorno depresivo?

10.4. En un estudio sobre tabaquismo se ha seleccionado una muestra aleatoria de 80 sujetos con problemas de tipo respiratorio. Por cada uno de estos sujetos se han seleccionado otros tres sin problemas respiratorios del mismo sexo, edad y nivel de estudios. Tras esto se ha clasificado a todos los sujetos como fumadores y no fumadores, obteniendo los resultados que muestra la siguiente tabla:

	Problemas respiratorios		
Tabaquismo	Sí	No	*Total*
Fumadores	52	60	112
No fumadores	28	180	208
Total	80	240	320

a. ¿Puede afirmarse que ser o no fumador está relacionado con tener o no problemas respiratorios? ($\alpha = 0{,}05$).
b. ¿Puede concluirse que el tabaco produce problemas respiratorios?

10.5. Se desea averiguar si utilizar o no el transporte público es independiente de la ciudad donde se vive. Para ello, a una muestra aleatoria de 10 personas de cada ciudad (4 ciudades) se les ha preguntado si utilizan o no el transporte público con regularidad. Las personas que han respondido afirmativamente en cada ciudad son las siguientes: 2 en Cuenca, 5 en Sevilla, 7 en Barcelona y 8 en Madrid.

a. ¿Qué prueba estadística debe aplicarse?
b. Plantea la hipótesis nula.
c. Si el punto crítico correspondiente a un nivel de significación de $\alpha = 0{,}05$ vale 7,81, ¿qué valores del estadístico del contraste llevarán a mantener H_0?
d. Si el estadístico del contraste toma el valor 8,48, ¿qué decisión debe tomarse sobre H_0 y por qué?
e. ¿Cuál es la conclusión del estudio?

10.6. Se desea contrastar la hipótesis de que una variable se distribuye según el modelo de probabilidad multinomial. Para tal fin se utiliza la prueba X^2 de Pearson y en una muestra aleatoria de tamaño n se obtiene $X^2 = 2{,}17$. Utilizando un nivel de significación de 0,05 y sabiendo que $P\,(X^2 \geq 2{,}17) = 0{,}96$:

a. ¿Debe rechazarse H_0? ¿Por qué?
b. ¿A partir de qué nivel de significación podría rechazarse H_0?

10.7. En un contraste sobre igualdad de proporciones se ha obtenido para el estadístico de contraste el valor 3. Sabiendo que $P\,(X^2 \leq 3) = 0{,}70$ y utilizando un nivel de significación de 0,05:

a. ¿Qué decisión debe tomarse sobre H_0? ¿Por qué?
b. ¿A partir de qué nivel de significación se puede empezar a rechazar H_0?

10.8. Se seleccionan dos muestras aleatorias de estudiantes de psicología: una de primer curso y otra de segundo curso. Se pregunta a los estudiantes si prefieren examen sólo teórico, sólo práctico o ambos. Los datos se analizan con la prueba X^2 de Pearson. Plantear las correspondientes hipótesis nula y alternativa tanto en términos estadísticos como de investigación.

10.9. Consideremos que en el estudio del ejercicio anterior se obtiene, para el estadístico del contraste, un valor $X^2 = 3{,}27$ tal que $P\,(X^2 \geq 3{,}27) = 0{,}0001$. Lo razonable será concluir que:

a. La preferencia por uno u otro tipo de examen depende del curso.
b. La preferencia por uno u otro tipo de examen no es la misma en las dos muestras.

c. La proporción de estudiantes que prefiere cada tipo de examen es la misma en los dos cursos.
d. La proporción de estudiantes que prefiere cada tipo de examen no es la misma en los dos cursos.
e. Todas las anteriores alternativas son incorrectas.

10.10. La hipótesis nula que puede contrastar la prueba X^2 de Pearson admite varias formulaciones. Señalar, entre las siguientes, la(s) correcta(s):

a. Las dos muestras son homogéneas.
b. Las J muestras son linealmente independientes.
c. Los parámetros siguen el modelo binomial.
d. Las J poblaciones tienen la misma distribución.
e. Las variables X e Y son independientes.

Soluciones

10.1. *a*. 27,65% (553 de 2.000).
b. 16,05% (321 de 2000).
c. 7,18% (31 de 432).
d. 23,08% (51 de 221).
e. Prueba X^2 de Pearson.

1. *Hipótesis*: H_0: la causa de la muerte es independiente de la edad.
2. *Supuestos*: muestra aleatoria de 2.000 observaciones independientemente clasificadas en las 18 categorías resultantes de combinar dos variables categóricas.
3. *Estadístico del contraste*. Para obtener el estadístico de Pearson (ecuación [10.7]) es necesario calcular primero las frecuencias esperadas. La siguiente tabla recoge estas frecuencias; se han obtenido aplicando la ecuación [10.6]. Por ejemplo, la frecuencia esperada correspondiente a la primera casilla de la tabla se ha obtenido así: $\hat{m}_{11} = 221\,(112)/2.000 = 12{,}38$.

	15-30 años	31-50 años	51-70 años
Accidente	12,38	47,74	160,89
Cáncer	30,97	119,45	402,58
Enfermedad cardíaca	21,22	81,86	275,91
Homicidio	2,80	10,80	36,40
Suicidio	4,59	17,71	59,70
Otra causa	40,04	154,44	520,52

$$X^2 = \frac{(51-12{,}38)^2}{12{,}38} + \frac{(93-47{,}74)^2}{47{,}74} + \cdots + \frac{(715-520{,}52)^2}{520{,}52} = 410{,}03\,.$$

4. *Distribución muestral*: χ^2 con $(I-1)(J-1) = (6-1)(3-1) = 10$ grados de libertad (χ^2_{10}).
5. *Zona crítica*: $X^2 \geq \chi^2_{10;\,0{,}95} = 18{,}31$ (ver Tabla D del Apéndice final).
6. *Decisión*: el valor del estadístico del contraste (410,03) es mayor que el punto crítico (18,31); por tanto, se rechaza H_0 y se concluye que la causa de la muerte está relacionada con la edad.

f. La intensidad de la relación puede estimarse utilizando alguna medida de asociación como, por ejemplo, el coeficiente de contingencia (ver ecuación [10.9]): $C = \sqrt{410{,}03/(410{,}03 + n)} = 0{,}41$. Este valor puede parecer, en principio, que refleja una asociación de intensidad media. Sin embargo, teniendo en cuenta que la interpretación de una medida de asociación depende del contexto, es muy probable que con este tipo de variables (causa de la muerte, grupos de edad) y con una muestra de 2.000 personas, un valor de 0,41 esté reflejando un grado de asociación bastante importante.

g. Para saber qué causas de muerte predominan en cada grupo de edad vamos a calcular los residuos tipificados corregidos. La siguiente tabla muestra el valor de estos residuos tras aplicar la ecuación [10.14]. Por ejemplo, el residuo tipificado corregido correspondiente a la primera casilla de la tabla se obtiene mediante: $Z_{R_{ij\,(\text{corregido})}} = (51-12{,}38)/\sqrt{12{,}38\,(1-221/2000)\,(1-112/2000)} = 11{,}98$.

	15-30 años	31-50 años	51-70 años
Accidente	11,98	7,85	−13,45
Cáncer	−5,43	−6,74	9,03
Enfermedad cardíaca	−4,27	−3,86	5,78
Homicidio	6,35	4,25	−7,21
Suicidio	4,61	3,64	−5,75
Otra causa	−3,25	1,42	0,36

Los residuos tipificados mayores que 1,96 (cuantil 97,5 de la distribución normal tipificada) permiten identificar las casillas en las que el número de casos es significativamente mayor que el que pronostica la hipótesis de independencia (es decir, significativamente mayor que el que cabe esperar por azar cuando las variables son realmente independientes).

Por tanto, los residuos mayores que 1,96 están indicando qué causas de muerte predominan en cada grupo de edad. Y la pauta que se observa en el ejemplo es bastante clara: en los dos grupos de menor edad (es decir, 15-30 y 31-50) hay más muertes de las esperadas por accidente, homicidio y suicidio; mientras que en el grupo de mayor edad (es decir, 51-70) hay más muertes de las esperadas por cáncer y enfermedad cardíaca.

10.2. *a*. Porcentajes de fila (la suma de los porcentajes de cada fila suman 100).

	A favor	En contra	*Total*
Solteros	86,7 %	13,3 %	100 %
Casados	25,0 %	75,0 %	100 %
Divorciados / separados	40,0 %	60,0 %	100 %
Viudos	40,0 %	60,0 %	100 %
Total	48,0 %	52,0 %	100 %

b. Prueba X^2 de Pearson.

1. *Hipótesis*: H_0: la actitud hacia el aborto es independiente del estado civil.
2. *Supuestos*: muestra aleatoria de 500 observaciones independientemente clasificadas en las 8 categorías resultantes de combinar dos variables categóricas.
3. *Estadístico del contraste*. Para obtener el estadístico de Pearson (ecuación 10.7) es necesario calcular primero las frecuencias esperadas. La siguiente tabla recoge estas frecuencias; se han obtenido aplicando la ecuación 10.6. Por ejemplo, la frecuencia esperada correspondiente a la primera casilla de la tabla se obtiene mediante: $\hat{m}_{11} = 150(240)/500 = 72$.

	A favor	En contra	*Total*
Solteros	72	78	150
Casados	96	104	200
Divorciados / separados	48	52	100
Viudos	24	26	50
Total	240	260	500

$$X^2 = \frac{(130-72)^2}{72} + \frac{(20-78)^2}{78} + \cdots + \frac{(30-26)^2}{26} = 136{,}08$$

4. *Distribución muestral*: χ^2 con $(I-1)(J-1) = (4-1)(2-1) = 3$ grados de libertad (χ^2_3).
5. *Zona crítica*: $X^2 \geq \chi^2_{3;\,0,95} = 7{,}81$ (ver Tabla D del Apéndice final).
6. *Decisión*: como el valor del estadístico del contraste (136,08) es mayor que el punto crítico (7,81), se rechaza H_0. Por tanto, puede concluirse que la actitud hacia el aborto está relacionada con el estado civil.

c. Para interpretar la relación encontrada debemos basarnos en los porcentajes de fila o columna (utilizaremos los porcentajes de fila que ya tenemos calculados en el apartado *a*) y los residuos tipificados corregidos. La siguiente tabla ofrece el valor que toman estos residuos tras aplicar la ecuación [10.14]. Por ejemplo, el residuo tipificado corregido correspondiente a la primera casilla de la tabla se obtiene así: $Z_{R_{ij\,(\text{corregido})}} = (130-72)/\sqrt{72\,(1-150/500)\,(1-240/500)} = 11{,}33$.

	A favor	En contra
Solteros	11,3	−11,3
Casados	−8,4	8,4
Divorciados / separados	−1,8	1,8
Viudos	−1,2	1,2

Los porcentajes de fila calculados en el apartado *a* de este mismo ejercicio indican que el porcentaje de personas que está a favor del aborto (48%) es muy parecido al de personas que están en contra (52%). Pero el hecho de que las variables estén relacionadas significa que esa pauta en la actitud hacia el aborto (48% frente a 52%) no se reproduce en todos los estados civiles. Los residuos tipificados corregidos permiten precisar dónde se rompe esa pauta. En concreto, entre los *solteros* se observa un desplazamiento significativo de casos hacia la categoría *a favor* (pues 11,3 es mayor que 1,96), mientras que entre los *casados* se observa un desplazamiento significativo de casos hacia el *en contra* (pues 8,4 es mayor que 1,96). Es decir, hay más solteros a favor y más casados en contra de lo que pronostica la hipótesis de independencia. Los porcentajes de fila en los otros dos grupos no parecen alejarse de lo esperado (sus residuos están comprendidos entre −1,96 y 1,96).

10.3. Prueba X^2 de Pearson.

1. *Hipótesis*: H_0: el hábitat es independiente de padecer o no trastorno depresivo.
2. *Supuestos*: muestra aleatoria de 300 observaciones independientemente clasificadas en las 6 categorías resultantes de combinar dos variables categóricas.
3. *Estadístico del contraste*. Para obtener el estadístico de Pearson (ecuación [10.7]) comenzamos calculando las frecuencias esperadas. La siguiente tabla recoge estas frecuencias; se han obtenido aplicando la ecuación [10.6]. Por ejemplo, la frecuencia esperada correspondiente a la primera casilla de la tabla se obtiene mediante: $\hat{m}_{11} = 60(100)/300 = 20$.

	Rural	Semiurbano	Urbano	*Total*
Sí	20	20	20	60
No	80	80	80	240
Total	100	100	100	300

$$X^2 = \frac{(12-20)^2}{20} + \frac{(16-20)^2}{20} + \cdots + \frac{(68-80)^2}{80} = 14{,}0$$

4. *Distribución muestral*: χ^2 con $(I-1)(J-1) = (2-1)(3-1) = 2$ grados de libertad (χ^2_2).
5. *Zona crítica*: $X^2 \geq \chi^2_{2;\,0,95} = 5{,}99$ (ver Tabla D del Apéndice final).
6. *Decisión*: como el valor del estadístico del contraste (14,0) es mayor que el punto crítico (5,99), se rechaza H_0. Por tanto, puede concluirse que, en la población de desempleados, las variables *hábitat* y *padecer o no trastorno depresivo* están relacionadas. O, lo que es lo mismo: la proporción de desempleados con trastorno depresivo no es la misma en los tres hábitats estudiados.

10.4. *a*. Prueba X^2 de Pearson.

1. *Hipótesis*: H_0: el tabaquismo es independiente de los problemas respiratorios.
2. *Supuestos*: muestra aleatoria de 320 observaciones independientemente clasificadas en las 4 categorías resultantes de combinar dos variables categóricas.
3. *Estadístico del contraste*. Para obtener el estadístico de Pearson (ecuación [10.7]) comenzamos calculando las frecuencias esperadas. La siguiente tabla recoge estas frecuencias; se han obtenido aplicando la ecuación [10.6]. Por ejemplo, la frecuencia esperada correspondiente a la primera casilla de la tabla se obtiene así: $\hat{m}_{11} = 112(80)/320 = 28$.

	Sí	No
Fumadores	28	84
No fumadores	52	156

$$X^2 = \frac{(52-28)^2}{28} + \frac{(60-84)^2}{84} + \frac{(28-52)^2}{52} + \frac{(180-156)^2}{156} = 42{,}20$$

4. *Distribución muestral*: χ^2 con $(I-1)(J-1) = (2-1)(2-1) = 1$ grados de libertad (χ^2_1).
5. *Zona crítica*: $X^2 \geq \chi^2_{1;\,0,95} = 3{,}84$ (ver Tabla D del Apéndice final).
6. *Decisión*: como el valor del estadístico del contraste (42,20) es mayor que el punto crítico (3,84), se rechaza H_0. Por tanto, puede concluirse que el tabaquismo está relacionado con tener o no problemas respiratorios. O, lo que es lo mismo: la proporción de sujetos con problemas respiratorios no es la misma entre los fumadores y entre los no fumadores.

b. Con un diseño de estas características (no hay asignación aleatoria de los sujetos a las condiciones del estudio) no es posible determinar si el tabaquismo produce problemas respiratorios. Sólo es posible afirmar que existe relación.

10.5. *a*. Prueba X^2 de Pearson sobre independencia o igualdad de proporciones.

b. H_0: usar o no transporte público es independiente de la ciudad donde se vive (o bien, la proporción de usuarios de transporte público es la misma en las 4 ciudades).

c. $X^2 < 7{,}81$.

d. Rechazarla, porque $9{,}25 > 7{,}81$.

e. Las proporción de personas que usa transporte público no es la misma en las cuatro ciudades. Es decir, usar o no transporte público está relacionado con la ciudad donde se vive.

10.6. *a*. No. Por encima (a la derecha) del valor $X^2 = 2{,}17$ queda un área de 0,96. Y puesto que en los contrastes con el estadístico X^2 la zona crítica siempre está situada a la derecha de su distribución muestral, el estadístico $X^2 = 2{,}17$ está en la zona de aceptación (muy alejado de la zona crítica; en el otro extremo de la distribución).
b. El nivel de significación a partir del cual podría rechazarse la hipótesis nula es justamente el nivel crítico, es decir, 0,96.

10.7. *a*. Mantenerla; porque $P(X^2 \geq 3) = 0{,}30 > 0{,}05$.
b. $p = 0{,}30$.

10.8. La situación descrita exige relacionar dos variables categóricas: curso y tipo de examen; o lo que es lo mismo, comparar dos muestras independientes (una de estudiantes de primero y otra de estudiantes de segundo) en una variable categórica (tipo de examen).

La hipótesis de investigación debe estar referida a las distribuciones de ambas poblaciones en la variable categórica. Podría formularse en los siguientes términos: *los estudiantes de primero difieren de los estudiantes de segundo en el tipo de examen que prefieren*. Las hipótesis nula y alternativa que corresponden a esta hipótesis de investigación son las siguientes:

- H_0: las variables *curso* y *tipo de examen* son independientes.
- H_1: las variables *curso* y *tipo de examen* no son independientes.

10.9. Puesto que el nivel crítico (0,0001) asociado al estadístico del contraste es muy pequeño (menor que cualquier nivel de significación estándar), la decisión correcta será rechazar H_0 y concluir que los estudiantes de 1º y 2º difieren en el tipo de examen que prefieren. Esto nos debe llevar a seleccionar la alternativa *d* como correcta y descartar las alternativas *c* y *e*. La alternativa *a* es incorrecta porque se insinúa que la relación encontrada es de naturaleza causal ("la preferencia... *depende* del curso") que no puede afirmarse en un estudio de estas características. Y la alternativa *b* es incorrecta porque las conclusiones de un contraste no se refieren a las muestras utilizadas para realizar el contraste, sino a las poblaciones a las que esas muestras representan.

10.10. Las conclusiones de un contraste nunca se refieren a las muestras utilizadas en el contraste, sino a las poblaciones a las que representan; esto permite descartar las dos primeras alternativas. Los parámetros no son *variables*, sino *constantes*; y las constantes no tienen distribución de probabilidad; esto descarta la alternativa *c*. Las alternativas *d* y *e* son correctas: la primera se refiere a la hipótesis de *igualdad de proporciones*; la segunda, a la hipótesis de *independencia* entre variables.

11

Inferencia con una variable categórica y una cuantitativa

Trabajar simultáneamente con una variable categórica y una cuantitativa significa, por lo general, trabajar con una variable que define grupos (la categórica) y una variable en la cual se desea comparar los grupos (la cuantitativa). Si la variable categórica tiene sólo dos categorías y, por tanto, define sólo dos grupos, lo habitual es compararlos mediante la prueba *T* de *Student para muestras independientes*; si la variable categórica tiene más de dos categorías y, por tanto, define más de dos grupos, lo habitual es compararlos mediante el *análisis de varianza de un factor*. Este capítulo se centra en la prueba *T*; el análisis de varianza se estudia en el segundo volumen[1].

Al analizar una variable categórica y una cuantitativa, *comparar* y *relacionar* significan la misma cosa: si los grupos definidos por la variable categórica difieren en la variable cuantitativa, entonces la variable categórica y la cuantitativa están relacionadas. Por tanto, la misma herramienta que sirve para constatar la presencia de diferencias sirve también para constatar la presencia de relación (recordemos que esto también ocurría en el caso de dos variables categóricas).

La prueba *T* de *Student* para muestras independientes, o *contraste sobre dos medias independientes*[2], es una técnica de análisis frecuentemente utilizada para analizar datos. Sirve para averiguar si dos grupos difieren en alguna variable cuantitativa de interés. Es decir, sirve, por ejemplo, para comparar en una variable cuantitativa un grupo experimental y un grupo control, o dos tratamientos, o dos colectivos distintos (hombres y mujeres; fumadores y no fumadores; etc.).

---

[1] Tanto en el caso de dos grupos como en el de más de dos existen algunos procedimientos alternativos, generalmente agrupados bajo la denominación de *no paramétricos*, que también se tratan en el segundo volumen.

[2] En este contexto, muestras *independientes* es sinónimo de grupos aleatorios. Esto implica que se está trabajando con dos grupos de sujetos distintos, aleatoriamente seleccionados de sus respectivas poblaciones. En el siguiente capítulo se estudia la prueba *T* de *Student* para muestras *relacionadas*. El significado de muestras relacionadas se contrapone al de muestras independientes: no se trata de dos grupos aleatorios, sino de un sólo grupo de sujetos al que se toman dos medidas o una medida en dos momentos distintos (ver, en el siguiente capítulo, el apartado *Muestras relacionadas*).

Bajo ciertas condiciones, esta herramienta estadística es la idónea para comparar las medias de dos grupos de puntuaciones. Y esto, independientemente de que el diseño utilizado para recoger los datos sea observacional, correlacional o experimental. Esta idea es especialmente importante. La prueba *T* permite averiguar si dos grupos difieren; sólo eso; no permite ir más allá. Para poder afirmar que las diferencias encontradas reflejan una relación de naturaleza causal es necesario atenerse a las características del diseño. Y ya hemos señalado (ver, en el Capítulo 1, el apartado *Niveles de indagación*) que sólo en el contexto de los diseños experimentales es posible hacer afirmaciones sobre la naturaleza causal de una relación (al margen, por supuesto, de la posible existencia de alguna teoría bien fundamentada de la que se deriven tales afirmaciones).

La prueba *T* de Student para muestras independientes

Nos encontramos en el siguiente escenario: dos poblaciones Y_1 e Y_2, con medias μ_{Y_1} y μ_{Y_2}, y dos muestras de tamaños n_1 y n_2, seleccionadas aleatoria e independientemente de sus respectivas poblaciones. El objetivo del análisis es el de contrastar la hipótesis de que las medias poblacionales son iguales: $\mu_{Y_1} = \mu_{Y_2}$ (o lo que es lo mismo: $\mu_{Y_1} - \mu_{Y_2} = 0$).

¿Cómo hacer esto? Del mismo modo que la media muestral es el mejor estimador de la media poblacional, el mejor estimador de la diferencia entre dos medias poblacionales es justamente la diferencia entre las correspondientes medias muestrales $\bar{Y}_1 - \bar{Y}_2$. Ahora bien, para que un estadístico sirva para contrastar una hipótesis, no basta con que ofrezca información relevante sobre esa hipótesis, es necesario, además, que tenga distribución muestral conocida.

Tenemos dos poblaciones (ver Figura 11.1). Seleccionando una muestra aleatoria de cada población y calculando $\bar{Y}_1$ e $\bar{Y}_2$ podremos obtener la diferencia $\bar{Y}_1 - \bar{Y}_2$. El valor de esta diferencia dependerá, obviamente, de las muestras concretas seleccionadas. Repitiendo el proceso una vez más obtendremos un nuevo valor para $\bar{Y}_1 - \bar{Y}_2$ que será, probablemente, diferente del anterior. Y repitiendo el proceso un número indefinido de veces podremos conocer todos los posibles valores de $\bar{Y}_1 - \bar{Y}_2$ y la frecuencia con la que se repite cada uno de ellos; es decir, podremos conocer la *distribución muestral* del estadístico $\bar{Y}_1 - \bar{Y}_2$ (ver Figura 11.2).

¿Cuáles son las características de esta distribución muestral? Como $\bar{Y}_1 - \bar{Y}_2$ es el resultado de combinar dos variables aleatorias independientes entre sí, se verifica que[3]

$$\begin{aligned} E(\bar{Y}_1 - \bar{Y}_2) &= E(\bar{Y}_1) - E(\bar{Y}_2) = \mu_{Y_1} - \mu_{Y_2} \\ V(\bar{Y}_1 - \bar{Y}_2) &= \sigma^2_{\bar{Y}_1 - \bar{Y}_2} = \sigma^2_{\bar{Y}_1} + \sigma^2_{\bar{Y}_2} = \sigma^2_{Y_1}/n_1 + \sigma^2_{Y_2}/n_2 \end{aligned} \qquad [11.1]$$

Si las dos poblaciones muestreadas son normales, también serán normales las distribuciones muestrales de $\bar{Y}_1$ y de $\bar{Y}_2$ (ver Capítulo 6); y puesto que $\bar{Y}_1 - \bar{Y}_2$ es una combinación lineal de $\bar{Y}_1$ e $\bar{Y}_2$, también su distribución muestral será normal. Además, sabemos por el teorema del límite central que, cualquiera que sea la forma de las poblacionales originales, las distribuciones muestrales de $\bar{Y}_1$ y de $\bar{Y}_2$ tienden a la normal a medida que los tamaños muestrales van aumentando. Y lo mismo ocurre con $\bar{Y}_1 - \bar{Y}_2$. Por tanto, el estadístico $\bar{Y}_1 - \bar{Y}_2$, bajo las menciona-

[3] La varianza de la suma de dos variables es igual a la suma de las varianzas individuales más el doble de las covarianzas (el concepto de covarianza se estudia en el próximo capítulo). La varianza de la resta de dos variables es la suma de las varianzas individuales menos el doble de las covarianzas. Si las variables son independientes, las covarianzas valen cero y tanto la varianza de la suma como la varianza de la resta es igual a la suma de las varianzas individuales.

das circunstancias, se distribuye normalmente con el valor esperado y la varianza definidos en [11.1]. En la Figura 11.2 está representada la distribución muestral de $\bar{Y}_1 - \bar{Y}_2$ junto con su valor esperado y su error típico (raíz cuadrada de la varianza). Es importante no confundir las distribuciones originales (Figura 11.1) con la distribución muestral de $\bar{Y}_1 - \bar{Y}_2$ (Figura 11.2).

Figura 11.1. Dos poblaciones normales (Y_1, Y_2) con sus correspondientes medias y desviaciones típicas

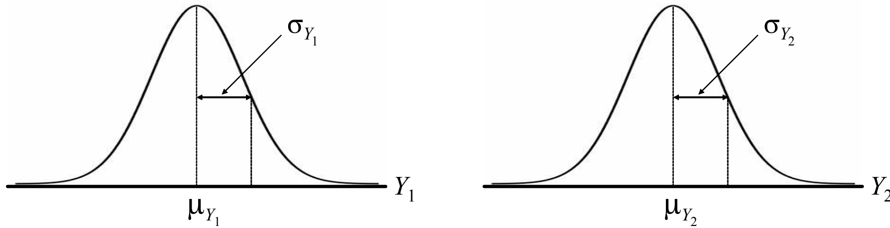
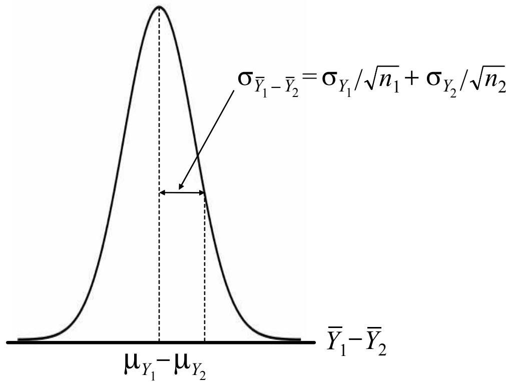


Figura 11.2. Distribución muestral de la diferencia entre dos medias independientes



Ahora bien, si la diferencia entre las medias muestrales ($\bar{Y}_1 - \bar{Y}_2$) se distribuye normalmente, entonces la transformación

$$Z = \frac{(\bar{Y}_1 - \bar{Y}_2) - E(\bar{Y}_1 - \bar{Y}_2)}{\sigma_{\bar{Y}_1 - \bar{Y}_2}} = \frac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_{Y_1} - \mu_{Y_2})}{\sqrt{\sigma^2_{Y_1}/n_1 + \sigma^2_{Y_2}/n_2}} \qquad [11.2]$$

se distribuye $N(0, 1)$. Por tanto, la transformación Z puede utilizarse, junto con la distribución $N(0, 1)$, para conocer las probabilidades asociadas al estadístico ($\bar{Y}_1 - \bar{Y}_2$) en su distribución muestral. Lo cual significa que tenemos todo lo necesario para poder contrastar hipótesis referidas al parámetro $\mu_1 - \mu_2$.

Pero la utilidad del estadístico Z propuesto en [11.2] es bastante escasa debido a que exige conocer el valor de las varianzas poblacionales. En el trabajo aplicado, raramente se dan situaciones en las que, siendo desconocidas las medias poblacionales (razón por la cual se efectúa un contraste de hipótesis sobre ellas), se conozcan las varianzas. Lo habitual es, más bien, que las varianzas poblacionales sean, al igual que las medias, desconocidas; en cuyo caso el error típico de la distribución muestral de $\bar{Y}_1 - \bar{Y}_2$ (el denominador de la ecuación [11.2]) será igualmente desconocido y habrá que estimarlo. Veamos cómo hacerlo.

Asumiendo varianzas iguales

Si se asume que las varianzas poblacionales son iguales (es decir, $\sigma^2_{Y_1} = \sigma^2_{Y_2} = \sigma^2_Y$), sólo será necesario estimar un parámetro (σ^2_Y). Y puesto que los dos estimadores de ese parámetro ($S^2_{Y_1}$ y $S^2_{Y_2}$) son independientes entre sí (recordemos que proceden de muestras independientes), lo razonable será combinar ambos para obtener una única estimación de σ^2_Y (porque lo más probable es que la combinación de ambos permita obtener una estimación de σ^2_Y más *exacta* que la de cada uno por separado). Esta estrategia conduce a

$$\hat{\sigma}^2_Y = \frac{(n_1 - 1)\, S^2_{Y_1} + (n_2 - 1)\, S^2_{Y_2}}{n_1 + n_2 - 2} \qquad [11.3]$$

como estimador insesgado de σ^2_Y. Sustituyendo ahora σ^2_Y en [11.1] por su estimador en [11.3] se obtiene

$$\hat{\sigma}^2_{\bar{Y}_1 - \bar{Y}_2} = \frac{\hat{\sigma}^2_Y}{n_1} + \frac{\hat{\sigma}^2_Y}{n_2} = \frac{(n_1 - 1)\, S^2_{Y_1} + (n_2 - 1)\, S^2_{Y_2}}{n_1 + n_2 - 2} \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \qquad [11.4]$$

como estimador insesgado de $\sigma^2_{\bar{Y}_1 - \bar{Y}_2}$. A partir de aquí es fácil demostrar (ver Apéndice 11) que la transformación

$$T = \frac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_{Y_1} - \mu_{Y_2})}{\hat{\sigma}_{\bar{Y}_1 - \bar{Y}_2}} \qquad [11.5]$$

ya no se distribuye $N(0, 1)$, sino según el modelo de probabilidad t de *Student* con $n_1 + n_2 - 2$ grados de libertad. Por tanto, es posible utilizar la transformación T propuesta en [11.5] junto con la distribución t para conocer las probabilidades asociadas al estadístico $\bar{Y}_1 - \bar{Y}_2$ cuando, desconociendo las varianzas poblacionales, se asume que son iguales y se estiman a partir de la combinación ponderada de $S^2_{Y_1}$ y $S^2_{Y_2}$. En consecuencia, tenemos todo lo necesario para poder diseñar un contraste hipótesis sobre el parámetro $\mu_1 - \mu_2$.

Además, conociendo la distribución muestral del estadístico $\bar{Y}_1 - \bar{Y}_2$ y su error típico, también es posible construir intervalos de confianza para el parámetro $\mu_1 - \mu_2$. Recordemos (ver Capítulo 7, ecuación [7.4]) que la estrategia utilizada para construir un intervalo de confianza consiste en sumar y restar al estadístico utilizado como estimador una cantidad que llamamos *error máximo*. En el caso de la diferencia entre dos medias tendremos:

$$IC_{|\mu_{Y_1} - \mu_{Y_2}|} = |\bar{Y}_1 - \bar{Y}_2| \pm E_{\text{máx}}$$

La cantidad $E_{\text{máx}}$ se obtiene a partir del error típico del estimador, intentando abarcar un área de la distribución muestral tal que la probabilidad de encontrar valores dentro de ella valga $1 - \alpha$. En consecuencia:

$$IC_{|\mu_{Y_1} - \mu_{Y_2}|} = |\bar{Y}_1 - \bar{Y}_2| \pm t_{n_1 + n_2 - 2;\, 1 - \alpha/2}\, \hat{\sigma}_{\bar{Y}_1 - \bar{Y}_2} \qquad [11.6]$$

El Cuadro 11.1 ofrece un resumen del contraste sobre dos medias independientes, incluyendo el cálculo del nivel crítico y del intervalo de confianza.

Cuadro 11.1. Resumen del contraste de hipótesis sobre dos medias independientes asumiendo varianzas poblacionales iguales (prueba *T* de *Student* para muestras independientes)

1. *Hipótesis*[4]:
   a. Contraste bilateral: H_0: $\mu_{Y_1} - \mu_{Y_2} = k_0$; H_1: $\mu_{Y_1} - \mu_{Y_2} \neq k_0$.
   b. Contraste unilateral derecho: H_0: $\mu_{Y_1} - \mu_{Y_2} \leq k_0$; H_1: $\mu_{Y_1} - \mu_{Y_2} > k_0$.
   c. Contraste unilateral izquierdo: H_0: $\mu_{Y_1} - \mu_{Y_2} \geq k_0$; H_1: $\mu_{Y_1} - \mu_{Y_2} < k_0$.
2. *Supuestos*: tenemos dos muestras de tamaños n_1 y n_2 seleccionadas aleatoria e independientemente de dos poblaciones normales cuyas varianzas, aunque desconocidas, se asumen iguales.
3. *Estadístico del contraste*[5] (basado en [11.4] y [11.5]:

$$T = \frac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_{Y_1} - \mu_{Y_2})}{\sqrt{\dfrac{(n_1-1)\,S^2_{Y_1} + (n_2-1)\,S^2_{Y_2}}{n_1+n_2-2}\left(\dfrac{1}{n_1}+\dfrac{1}{n_2}\right)}} \qquad [11.7]$$

4. *Distribución muestral*: T se distribuye según t_{gl} (con $gl = n_1 + n_2 - 2$).
5. *Zona crítica*:
   a. Contraste bilateral: $T \leq t_{gl;\,\alpha/2}$ y $T \geq t_{gl;\,1-\alpha/2}$.
   b. Contraste unilateral derecho: $T \geq t_{gl;\,1-\alpha}$.
   c. Contraste unilateral izquierdo: $T \leq t_{gl;\,\alpha}$.
6. *Regla de decisión*: rechazar H_0 si el estadístico T cae en la zona crítica; en caso contrario, mantenerla.
7. *Nivel crítico* (valor *p*):
   a. Contraste bilateral: $p = 2[P(T \geq |T_h|)]$, siendo T_h el valor concreto que toma el estadístico T.
   b. Contraste unilateral derecho: $p = P(T \geq T_h)$.
   c. Contraste unilateral izquierdo: $p = P(T \leq T_h)$.
8. *Intervalo de confianza* (ver [11.6]):

$$IC_{|\mu_{Y_1} - \mu_{Y_2}|} = |\bar{Y}_1 - \bar{Y}_2| \pm t_{n_1+n_2-2;\,1-\alpha/2}\;\hat{\sigma}_{\bar{Y}_1-\bar{Y}_2}$$

---

[4] Generalmente $k_0 = 0$, pues la hipótesis nula que habitualmente interesa contrastar es si las medias poblacionales son iguales. Por tanto, $\mu_{Y_1} - \mu_{Y_2} = k_0$ será, generalmente, $\mu_{Y_1} - \mu_{Y_2} = 0$, lo cual equivale a $\mu_{Y_1} = \mu_{Y_2}$.

[5] Puesto que, generalmente, la hipótesis nula afirmará $k_0 = 0$, la expresión $\mu_{Y_1} - \mu_{Y_2}$ desaparecerá del numerador del estadístico T. Esto será así incluso en los contrastes unilaterales con H_0: $\mu_{Y_1} \leq \mu_{Y_2}$ o H_0: $\mu_{Y_1} \geq \mu_{Y_2}$, pues, según se ha explicado ya en el Capítulo 8, el modelo estadístico en el que se basa un contraste se define a partir del signo "=" presente en H_0. Además, cuando los tamaños muestrales son iguales ($n_1 = n_2 = n$), el denominador de [11.7] se reduce a

$$\hat{\sigma}_{\bar{Y}_1-\bar{Y}_2} = \sqrt{(S^2_{Y_1} + S^2_{Y_2})/n} \qquad [11.8]$$

Ejemplo. La prueba T de Student para muestras independientes asumiendo varianzas poblacionales iguales

Un educador sospecha que los niños con problemas perceptivos aumentan su rendimiento, con entrenamiento adecuado, en preguntas del test Raven (Y) que habitualmente no resuelven por carecer de las estrategias adecuadas. Con el fin de obtener alguna evidencia sobre su sospecha ha seleccionado una muestra aleatoria de 20 niños con problemas perceptivos y los ha repartido, también aleatoriamente, en dos grupos. A un grupo (experimental) lo ha entrenado durante 2 meses en tareas de percepción de formas; el otro grupo no ha recibido entrenamiento (control). Terminado el entrenamiento, ha pasado a todos los sujetos el test Raven para obtener una medida del su rendimiento individual. La Tabla 11.1 muestra los resultados obtenidos. ¿Permiten estos datos afirmar que los sujetos entrenados en percepción de formas rinden mejor en el test Raven que los sujetos no entrenados? (α = 0,05).

Tabla 11.1. Resultados del test Raven

Grupos	*Sujetos*										n_j	$\bar{Y}_j$	S_{Y_j}
1 = Experimental	64	63	74	65	74	85	78	76	69	70	10	71,8	6,96
2 = Control	60	62	70	61	67	70	64	71	60	63	10	64,8	4,34

Tenemos una variable categórica (*grupo*) con dos niveles (1 = «experimental», 2 = «control») y una variable cuantitativa (Y = «puntuaciones en el test Raven») en la cual se desea comparar los grupos. Por tanto, tenemos una situación susceptible de ser analizada mediante la prueba T de *Student* para muestras independientes:

1. *Hipótesis*: H_0: $\mu_{Y_1} \leq \mu_{Y_2}$
   H_1: $\mu_{Y_1} > \mu_{Y_2}$

   Se está planteando un contraste unilateral derecho porque la expectativa del educador es que los sujetos entrenados (grupo experimental) mejoren sus puntuaciones en el test Raven, es decir, obtengan mejores puntuaciones que los sujetos no entrenados (grupo control).

2. *Supuestos*: asumimos que las puntuaciones en el test Raven se distribuyen normalmente en las dos poblaciones; desconocemos las varianzas poblacionales pero asumimos que son iguales; las muestras se han seleccionado de forma aleatoria e independientemente una de otra.

3. *Estadístico del contraste* (ver ecuaciones [11.7] y [11.8]):

$$T = \frac{(71{,}8 - 64{,}8) - 0}{\sqrt{\dfrac{9(6{,}96)^2 + 9(4{,}34)^2}{10 + 10 - 2}\left(\dfrac{1}{10} + \dfrac{1}{10}\right)}} = \frac{(71{,}8 - 64{,}8) - 0}{\sqrt{\dfrac{6{,}96^2 + 4{,}34^2}{10}}} = \frac{7}{2{,}593} = 2{,}70$$

4. *Distribución muestral*: T se distribuye según t con $gl = n_1 + n_2 - 2 = 10 + 10 - 2 = 18$.

5. *Zona crítica*: $T \geq t_{18;\,0,95} = 1{,}734$.
6. *Decisión*: como 2,70 > 1,734, se rechaza H_0. Se puede concluir que el promedio de los sujetos entrenados (grupo experimental) es significativamente más alto que el de los sujetos no entrenados (grupo control).
7. *Nivel crítico*: $p = P(T \geq 2{,}70) < 0{,}01$.
8. *Intervalo de confianza* ($\hat{\sigma}_{\bar{Y}_1-\bar{Y}_2} = 2{,}593$; denominador del estadístico T):

$$IC_{\mu_{Y_1}-\mu_{Y_2}} = |\bar{Y}_1 - \bar{Y}_2| \pm t_{gl;\,1-\alpha/2}\,\hat{\sigma}_{\bar{Y}_1-\bar{Y}_2} = |71{,}8 - 64{,}8| \pm t_{18;\,0{,}975}\,\hat{\sigma}_{\bar{Y}_1-\bar{Y}_2} =$$

$$= 7 \pm 2{,}101\,(2{,}573) = 7 \pm 5{,}45 = (1{,}55;\; 12{,}45)$$

Estimamos, con una confianza del 95%, que los sujetos que han recibido entrenamiento puntúan en el test Raven, en promedio, entre 1,55 y 12,45 puntos más que los que no han recibido entrenamiento (el hecho de que el intervalo de confianza sea poco preciso –aproximadamente 11 puntos de amplitud– se debe a que está calculado con muestras pequeñas; ya sabemos –ver Capítulo 7– que, simplemente incrementando el tamaño muestral, es posible aumentar la precisión del intervalo).

Independencia, normalidad e igualdad de varianzas

Para que el estadístico T se distribuya según el modelo de probabilidad t es necesario trabajar con muestras aleatoria e independientemente seleccionadas de poblaciones normales con varianzas iguales (recordemos que estas condiciones están en el origen de la obtención del estadístico). De estas condiciones iniciales es de donde se derivan los tres supuestos típicos del contraste sobre dos medias independientes: independencia entre las observaciones, normalidad de las poblaciones y homocedasticidad (igualdad entre las varianzas poblacionales).

En relación con el supuesto de *independencia* vale lo ya dicho en el capítulo anterior a propósito de la prueba T para una muestra. Únicamente hay que añadir que las observaciones de un grupo han de ser independientes, no sólo de las observaciones del otro grupo, sino del resto de observaciones de su mismo grupo.

Por lo que se refiere al supuesto de *normalidad*, su incumplimiento no acarrea consecuencias relevantes sobre las conclusiones del contraste[6] si los tamaños muestrales son razonablemente grandes. A este respecto, también sirve lo ya dicho en el capítulo anterior a propósito de la prueba T para una muestra; basta con sustituir “el tamaño muestral” por “la suma de los tamaños muestrales”. Además, si los tamaños muestrales son iguales, la prueba T para dos muestras independientes es más robusta que la prueba T para una muestra: si las dos poblaciones tienen forma parecida (aunque ésta sea asimétrica), tamaños muestrales tan pequeños como $n_1 = n_2 = 5$ permiten que la aproximación del estadístico T a la distribución t sea ya lo

[6] Ya sabemos que los *supuestos* de un contraste de hipótesis se refieren a las condiciones que deben darse para que el estadístico del contraste se ajuste bien a la distribución muestral propuesta. Cuando se dice que un estadístico se ve afectado por el incumplimiento de algún supuesto, o que el incumplimiento de algún supuesto tiene consecuencias no deseadas sobre las conclusiones del contraste, se quiere decir que ese estadístico ya no sigue exactamente la distribución muestral propuesta; lo cual se traduce en que la probabilidad de tomar una decisión equivocada cambia. Estudiaremos diferentes aspectos relacionados con esta problemática en el segundo volumen.

bastante buena (Posten, 1978). Si las distribuciones poblacionales no tienen la misma forma es necesario utilizar tamaños muestrales más grandes.

En lo relativo al supuesto de *homocedasticidad* (varianzas poblacionales iguales) las cosas no son tan favorables. Si los tamaños muestrales son iguales y el supuesto de normalidad no se incumple, el estadístico T es válido incluso con varianzas poblacionales muy diferentes (ver Ramsey, 1980). Pero si, aun siendo normales las poblaciones originales, los tamaños muestrales son muy distintos, asumir que las varianzas poblacionales son iguales puede conducir a conclusiones equivocadas (ver, por ejemplo, Boneau, 1960). En estos casos, lo apropiado es recurrir a la estrategia que se describe en el siguiente apartado.

Para decidir si es o no razonable asumir que las varianzas poblacionales son iguales, lo habitual es contrastar la hipótesis nula de igualdad de varianzas (este contraste se incluye en el Apéndice 11); si se rechaza esta hipótesis, no podrá asumirse que las varianzas poblacionales son iguales.

No asumiendo varianzas iguales

Si no puede asumirse que las varianzas poblacionales son iguales, no tiene sentido realizar una única estimación de las mismas a partir de la combinación ponderada de los dos estimadores disponibles ($S^2_{Y_1}$ y $S^2_{Y_2}$). Lo razonable será, más bien, utilizar cada uno de ellos como estimador de la varianza de su propia población (James, 1951; Welch, 1938), es decir:

$$\hat{\sigma}^2_{\bar{Y}_1-\bar{Y}_2} = \frac{\hat{\sigma}^2_Y}{n_1} + \frac{\hat{\sigma}^2_Y}{n_2} = \frac{S^2_{Y_1}}{n_1} + \frac{S^2_{Y_2}}{n_2} \qquad [11.9]$$

Con esta nueva estimación de la varianza de $\bar{Y}_1 - \bar{Y}_2$ es posible definir un nuevo estadístico al que llamaremos T' para distinguirlo de T:

$$T' = \frac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_{Y_1} - \mu_{Y_2})}{\hat{\sigma}_{\bar{Y}_1-\bar{Y}_2}} = \frac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_{Y_1} - \mu_{Y_2})}{\sqrt{S^2_{Y_1}/n_1 + S^2_{Y_2}/n_2}} \qquad [11.10]$$

La diferencia entre T y T' está únicamente en la forma de estimar el error típico del denominador. Pero justamente esa diferencia es la que genera un problema, pues es la responsable de que la distribución muestral de T' no sea la misma que la de T. No obstante, se trata de un problema poco importante porque contamos con algunos procedimientos que permiten conocer de forma aproximada la distribución muestral de T'.

Los primeros intentos de obtener la distribución de T' fueron iniciados por Behrens y continuados por Fisher (de ahí que el problema de la heterogeneidad de varianzas se conozca en estadística como el problema *Behrens-Fisher*). Pero las soluciones prácticas fueron aportadas por otros autores. Entre éstas, la de Welch (1938) es la que ha acaparado las preferencias de los investigadores[7] (es, además, la que incorpora el SPSS).

En la solución propuesta por Welch, el estadístico T' sigue siendo una variable distribuida según la t de Student, pero con un número desconocido de grados de libertad. La so-

[7] Al parecer, Satterthwaite (1946) llegó de forma independiente a la misma solución propuesta por Welch.

lución pasa por determinar los grados de libertad (gl') que corresponden a la distribución de T' mediante[8]

$$gl' = \frac{(S^2_{Y_1}/n_1 + S^2_{Y_2}/n_2)^2}{(S^2_{Y_1}/n_1)^2/(n_1-1) + (S^2_{Y_2}/n_2)^2/(n_2-1)} \qquad [11.11]$$

El resultado obtenido para gl' se redondea al entero más próximo. Se obtienen así unos grados de libertad comprendidos entre un mínimo y un máximo conocidos: el mínimo es el valor más pequeño de $n_1 - 1$ y $n_2 - 1$; el máximo es $n_1 + n_2 - 2$.

El estadístico T' y su distribución muestral (la distribución t con los grados de libertad corregidos) sirven para contrastar la hipótesis de igualdad de medias siguiendo los mismos pasos que en el contraste resumido en el Cuadro 11.1. Lo único que hay que hacer es sustituir T por T' y gl por gl'. Y para construir el intervalo de confianza hay que tener en cuenta que el error típico de la diferencia entre las medias muestrales ($\hat{\sigma}_{\bar{Y}_1 - \bar{Y}_2}$) ha cambiado: ya no se obtiene a partir de [11.4] sino a partir de [11.9].

Ejemplo. La prueba T de Student para muestras independientes no asumiendo varianzas poblacionales iguales

La Tabla 11.2 muestra un resumen de los datos obtenidos en un estudio sobre la relación entre el peso de los niños al nacer (Y) y el hecho de que la madre haya fumado o no durante la gestación. De los 122 niños incluidos en el estudio, 37 pertenecen a una muestra aleatoria de madres fumadoras y 85 a una muestra aleatoria de madres no fumadoras. ¿Permiten estos datos afirmar que el peso medio de los recién nacidos de madres fumadoras es menor que el peso medio de los recién nacidos de madres no fumadoras? ($\alpha = 0{,}05$).

Tabla 11.2. Descriptivos referidos al peso de 122 recién nacidos

Grupo	n_j	$\bar{Y}_j$	S_{Y_j}
1. Madres fumadoras	37	2,67	0,45
2. Madres no fumadoras	85	3,14	0,59

Tenemos una variable categórica (*grupo*) con dos niveles (1 = «madres fumadoras», 2 = «madres no fumadoras») y una variable cuantitativa (Y = *peso al nacer*) en la cual queremos comparar los grupos. Tenemos, por tanto, una situación susceptible de ser analizada mediante la prueba *T de Student* para muestras independientes:

1. *Hipótesis*: H_0: $\mu_{Y_1} \geq \mu_{Y_2}$
   H_1: $\mu_{Y_1} < \mu_{Y_2}$ (contraste unilateral izquierdo).

---

[8] El propio Welch (1947) ha propuesto una ligera modificación de [11.11] que podría ofrecer una solución más exacta para gl' (ver, por ejemplo, Pardo y San Martín, 1998, pág. 199). No obstante, la diferencia entre ambas soluciones suele ser más bien insignificante.

La hipótesis nula afirma que no hay diferencia en el peso medio de los recién nacidos de los dos grupos. La hipótesis alternativa afirma que el peso medio de los recién nacidos de madres fumadoras es menor que el de los recién nacidos de madres no fumadoras.

2. *Supuestos*: tenemos dos muestras aleatoria e independientemente seleccionadas de sus respectivas poblaciones. Los tamaños muestrales son lo bastante grandes como para pasar por alto el supuesto de normalidad. No asumimos varianzas poblacionales iguales.
3. *Estadístico del contraste* (ver ecuación [11.10]):

$$T' = \frac{(2{,}67 - 3{,}14) - 0}{\sqrt{0{,}45^2/37 + 0{,}59^2/85}} = \frac{-0{,}47}{0{,}098} = -4{,}80$$

4. *Distribución muestral*: T' se distribuye según $gl' \approx 89$ (aplicando la ecuación 11.11).

$$gl' = \frac{(0{,}45^2/37 + 0{,}59^2/85)^2}{(0{,}45^2/37)^2/36 + (0{,}59^2/85)^2/84} = 88{,}74$$

5. *Zona crítica* (contraste unilateral izquierdo): $T' \leq t_{89;\,0{,}05} = -1{,}662$.
6. *Decisión*: como $-4{,}80 < -1{,}662$, se rechaza H_0. Por tanto, puede afirmarse que los promedios poblacionales comparados son distintos: el peso medio de los recién nacidos de madres fumadoras es menor que el de los recién nacidos de madres no fumadoras.
7. *Nivel crítico*: $p = P(T' \leq -4{,}80) < 0{,}005$.
8. *Intervalo de confianza* ($\hat{\sigma}_{\bar{Y}_1 - \bar{Y}_2} = 0{,}098$; denominador del estadístico T'):

$$IC_{\mu_{Y_1} - \mu_{Y_2}} = |\bar{Y}_1 - \bar{Y}_2| \pm t_{gl;\,1-\alpha/2}\,\hat{\sigma}_{\bar{Y}_1 - \bar{Y}_2} = |2{,}67 - 3{,}14| \pm t_{89;\,0{,}975}\,\hat{\sigma}_{\bar{Y}_1 - \bar{Y}_2} =$$
$$= 0{,}47 \pm 1{,}986(0{,}098) = 0{,}47 \pm 0{,}19 = (0{,}28;\ 0{,}66)$$

No sabemos cuál es, en la población, la diferencia exacta entre el peso medio de los recién nacidos de madres fumadoras y de madres no fumadoras, pero estimamos, con una confianza del 95%, que se encuentra entre 280 y 660 gramos.

La prueba *T* de Student para muestras independientes con SPSS

El SPSS incluye, además de los dos estadísticos *T* estudiados en este capítulo (*T* para cuando puede asumirse que las varianzas poblacionales son iguales y T' para cuando no puede asumirse que son iguales), la prueba de *Levene* sobre igualdad de varianzas, la cual permite decidir si es o no razonable asumir varianzas poblacionales iguales y, consecuentemente, si debe utilizarse una u otra versión del estadístico *T*.

Estos contrastes están disponibles en la opción **Comparar medias > Prueba *T* para muestras independientes** del menú **Analizar**. La lista de variables del cuadro de diálogo principal contiene un listado con las variables numéricas y de cadena corta del archivo de datos.

Para llevar a cabo un contraste con las especificaciones que el procedimiento tiene establecidas por defecto:

1. Seleccionar la variable (o variables) cuantitativa en la cual se desea comparar los grupos y trasladarla a la lista **Contrastar variables**. A esta lista únicamente pueden trasladarse variables con formato numérico. Cada variable seleccionada genera un contraste, y cada contraste incluye, además de algunos estadísticos descriptivos, las dos versiones del estadístico (T y T') junto con sus correspondientes niveles críticos (valores p) e intervalos de confianza.

2. Seleccionar la variable categórica que define los dos grupos que se desea comparar y trasladarla al cuadro **Variable de agrupación**. Esta variable puede tener formato numérico o de cadena corta.

   Tras seleccionar la variable de agrupación, aparecen junto a ella dos interrogantes entre paréntesis que avisan de la necesidad de informar de los códigos (valores) que identifican a los dos grupos que se desea comparar (es necesario indicar estos códigos incluso aunque la variable de agrupación seleccionada sólo tenga dos valores). Para definir estos códigos, pulsar el botón **Definir grupos**.

   Si la variable de agrupación es categórica, los códigos que definen los dos grupos que se desea comparar deben introducirse en los cuadros de texto **Grupo 1** y **Grupo 2** de la opción **Usar valores especificados**. Los casos que posean otros códigos serán excluidos del análisis. Si la variable de agrupación es una variable con formato de cadena, es muy importante tener en cuenta que los códigos deben introducirse de forma exacta (vigilando la presencia de espacios, las mayúsculas y minúsculas, etc.).

   Si se desea utilizar como variable de agrupación una variable cuantitativa, la opción **Punto de corte** permite introducir un valor como punto de corte para definir dos grupos: los casos con puntuación igual o mayor que el punto de corte forman un grupo; el resto de los casos forman el otro grupo. Esta opción no está disponible si se elige una variable con formato de cadena como variable de agrupación.

El botón **Opciones** del cuadro de diálogo principal ofrece la posibilidad de controlar algunos aspectos del análisis; en concreto, el nivel de confianza con el que se desea construir el intervalo de confianza y el tratamiento que se desea dar a los valores perdidos.

La opción **Intervalo de confianza: *k*%** permite establecer, en escala porcentual, el *nivel de confianza* $(1-\alpha)$ con el que se desea obtener el intervalo de confianza para la diferencia de medias. El valor de k es, por defecto, 95, pero es posible introducir cualquier otro valor comprendido entre 0,01 y 99,99.

Las opciones del recuadro **Valores perdidos** permiten decidir qué tratamiento se desea dar a los casos con valor perdido. La opción **Excluir casos según análisis** excluye de cada análisis (de cada prueba T solicitada) los casos con valor perdido en la variable de agrupación o en la variable que se está contrastando en ese análisis; la opción **Excluir casos según lista** excluye de todos los análisis (de todas las pruebas T solicitadas) los casos con algún valor perdido en la variable de agrupación o en una cualquiera de las variables incluidas en la lista **Contrastar variables**.

Ejemplo. La prueba T de Student para muestras independientes con SPSS

Este ejemplo muestra cómo contrastar hipótesis sobre dos medias independientes utilizando el procedimiento **Prueba *T* para muestras independientes**. Se basa en los datos de la Tabla 11.1,

los cuales, recordemos, corresponden a una muestra de 20 sujetos con problemas perceptivos, la mitad de los cuales ha recibido entrenamiento para mejorar sus puntuaciones en el test Raven. Antes de poder llevar a cabo el análisis es necesario reproducir los datos de la tabla en el *Editor de datos*.

La Tabla 11.3 reproduce los datos de la Tabla 11.1. Y la Figura 11.3 ilustra cómo deben introducirse esos datos en el *Editor de datos* del SPSS. Se han creado dos variables: *grupo* y *raven* (puntuaciones en el test Raven). Los códigos 1 y 2 asignados a la variable *grupo* identifican los dos niveles de la variable. A estos códigos se les han asignado las siguientes etiquetas: 1 = «Experimental», 2 = «Control»). Nótese que la forma de introducir los datos refleja con exactitud la situación que intentamos analizar: dos variables (una categórica y una cuantitativa) y 20 casos.

Tabla 11.3. Datos de la Tabla 11.1

Grupo	
Experimental	Control
64	60
63	62
74	70
65	61
74	67
85	70
78	64
76	71
69	60
70	63

Figura 11.3. Datos de la Tabla 11.3

	grupo	raven
1	1	64
2	1	63
3	1	74
4	1	65
5	1	74
6	1	85
7	1	78
8	1	76
9	1	69
10	1	70
11	2	60
12	2	62
13	2	70
14	2	61
15	2	67
16	2	70
17	2	64
18	2	71
19	2	60
20	2	63

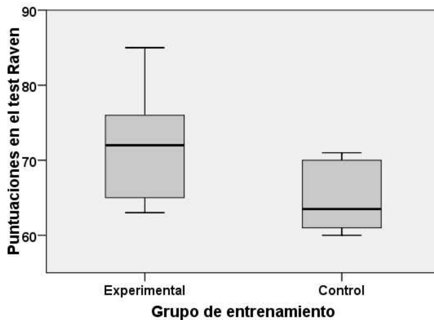
Nuestro interés principal es el de comparar los promedios de ambos grupos en el test Raven. Y esto es lo que haremos enseguida aplicando la prueba *T* de Student para muestras independientes. Pero antes, debemos formarnos una idea lo más exacta posible sobre las características de los datos que vamos a analizar. Esto, ya lo sabemos, se consigue aplicando herramientas descriptivas. Entre ellas, los diagramas de caja contienen toda la información relevante sobre una variable cuantitativa, pues no sólo informan sobre las tres características básicas de una distribución de puntuaciones (centro, dispersión y forma de la distribución), sino que permiten detectar la posible presencia de casos anómalos. Para obtener los diagramas de cajas correspondientes a cada grupo:

➤ Seleccionar la opción **Generador de gráficos** del menú **Gráficos** y, en la pestaña **Galería**, seleccionar la opción **Diagrama de caja** y, a la derecha, seleccionar el primero de los gráficos propuestos.

➤ Arrastrar la variable *grupo* al eje horizontal y la variable *raven* al eje vertical (para poder llevar la variable *grupo* al eje horizontal es imprescindible que tenga asignado un nivel de medida nominal u ordinal).

Aceptando estas selecciones se obtiene el diagrama de cajas que muestra la Figura 11.4. Para interpretar este gráfico prestamos atención a cuatro detalles: el centro (línea horizontal del interior de las cajas), la dispersión (altura de las caja y de los bigotes), la forma de la distribución (longitud de los bigotes, ubicación de los casos atípicos) y presencia de casos atípicos. Las cajas del ejemplo indican que el centro (mediana) del grupo experimental es mayor que el del grupo control; que la dispersión del grupo *experimental* es mayor que la del control, aunque en ambos casos se trata de distribuciones poco dispersas (ningún bigote se extiende hasta su límite); en ninguno de los dos grupos se observan asimetrías evidentes; y tampoco se observan casos atípicos.

Figura 11.4. Diagramas de caja correspondientes a los datos de la Tabla 11.3



Así pues, los promedios observados son distintos. La cuestión que hay que resolver ahora es si son lo bastante distintos como para poder afirmar que los correspondientes promedios poblacionales también lo son. Para ello:

➤ Seleccionar la opción **Comparar medias** > **Prueba *T* para muestras independientes** del menú **Analizar** y, en el cuadro de diálogo principal, trasladar la variable *raven* a la lista **Contrastar variables** y la variable *grupo* al cuadro de selección **Variable de agrupación**.

➤ Pulsar el botón **Definir grupos** e introducir los códigos 1 y 2 en los cuadros de texto **Grupo 1** y **Grupo 2**. Pulsar el botón **Continuar** para volver al cuadro de diálogo principal.

Aceptando estas selecciones, el *Visor* ofrece los resultados que muestran las Tablas 11.4 y 11.5. La Tabla 11.4 contiene información descriptiva sobre la variable cuantitativa (puntuaciones en el test Raven) en cada grupo: el número de casos válidos, la media, la desviación típica y el error típico de la media (el error típico de la media de cada grupo se obtiene dividiendo su desviación típica por la raíz cuadrada de su tamaño muestral).

Tabla 11.4. Estadísticos descriptivos del procedimiento *Prueba T para muestras independientes*

Puntuaciones en el test Raven

Grupo	N	Media	Desviación típica	Error típ. de la media
Experimental	10	71,80	6,957	2,200
Control	10	64,80	4,341	1,373

La Tabla 11.5 ofrece, en primer lugar, un contraste de la hipótesis de igualdad de varianzas basado en el estadístico *F* de Levene (ver Apéndice 11). El resultado de este contraste sirve para decidir si puede o no asumirse que las varianzas poblacionales son iguales: si la probabilidad asociada al estadístico de Levene (*sig.*) es menor que $\alpha = 0{,}05$, habrá que rechazar la hipótesis de igualdad de varianzas y, consecuentemente, no podrá asumirse que son iguales.

A continuación aparece el estadístico *T* (*t*), sus grados de libertad (*gl*), el nivel crítico bilateral o valor *p* (*sig.*; el nivel crítico unilateral se obtiene dividiendo por 2 el bilateral), la diferencia observada entre las medias de ambos grupos, el error típico de esa diferencia, y los límites del intervalo de confianza para la diferencia entre las medias poblacionales (calculado al 95%). Todo esto está calculado tanto para el caso de varianzas poblacionales iguales (línea encabezada *asumiendo varianzas iguales*) como para el caso de varianzas poblacionales distintas (línea encabezada *no asumiendo varianzas iguales*).

Tabla 11.5. Resumen del procedimiento *Prueba T para muestras independientes*

Puntuaciones en el test Raven

	Prueba de Levene para la igualdad de varianzas		Prueba T para la igualdad de medias						
								95% Intervalo de confianza para la diferencia	
	F	Sig.	t	gl	Sig. (bilateral)	Diferencia de medias	Error típ. de la diferencia	Inferior	Superior
Se han asumido varianzas iguales	2,028	,171	2,699	18	,015	7,000	2,593	1,552	12,448
No se han asumido varianzas iguales			2,699	15,086	,016	7,000	2,593	1,476	12,524

En el ejemplo, la probabilidad asociada al estadístico de Levene (*sig.* = 0,145 > 0,05) no permite rechazar la hipótesis de igualdad de varianzas[9] y, por tanto, puede asumirse que las varianzas poblacionales son iguales. En consecuencia, para contrastar la hipótesis de igualdad de medias utilizaremos los resultados de la fila encabezada *asumiendo varianzas iguales*: el estadístico *t* vale 2,699 y tiene asociado un nivel crítico bilateral de 0,015 (el unilateral vale la mitad). Puesto que el nivel crítico (0,015/2 = 0,0075; no olvidar que estamos planteando un contraste unilateral) es menor que 0,05, se puede rechazar la hipótesis nula de igualdad de medias y concluir que la media en el test Raven de los sujetos entrenados es mayor que la de los sujetos no entrenados.

No debe olvidarse que esta afirmación se refiere a las medias poblacionales (para ver si las medias muestrales son distintas no es necesario hacer ningún contraste). Y puesto que estamos afirmando que las medias poblacionales no son iguales, lo interesante es poder estimar cuánto vale la diferencia. Precisamente para esto es para lo que sirve el intervalo de confianza: podemos estimar, con una confianza del 95%, que la diferencia entre las medias poblacionales de ambos grupos se encuentra entre 1,55 y 12,45 puntos. El hecho de que el valor cero no se encuentre entre los límites del intervalo también permite rechazar la hipótesis de igualdad de medias.

[9] Al hacer este mismo ejemplo aplicando la ecuación [11.7] hemos optado por asumir varianzas poblacionales iguales (lo cual no suele ser un problema cuando los tamaños muestrales son iguales). El resultado de la prueba de Levene indica que, con estos datos, efectivamente no hay problema en asumir varianzas poblacionales iguales.

Apéndice 11

La distribución muestral del estadístico *T* asumiendo $\sigma_1 = \sigma_2$

El contraste sobre dos medias independientes propuesto en el Cuadro 11.1 se basa en la distribución muestral del estadístico $\bar{Y}_1 - \bar{Y}_2$ y en la transformación de ese estadístico en T. La distribución muestral de $\bar{Y}_1 - \bar{Y}_2$ sabemos que es normal porque es combinación lineal de dos variables que sabemos que se distribuyen normalmente. Pero, ¿cuál es la distribución muestral de la transformación T?

Al estudiar la distribución muestral de la varianza (ver, en el Apéndice 6, la ecuación [6.11]) hemos visto que

$$(n-1)\, S_Y^2 / \sigma_Y^2 \;=\; \chi^2_{n-1}$$

En consecuencia,

$$(n_1 - 1)\, S_{Y_1}^2 / \sigma_{Y_1}^2 \;=\; \chi^2_{n_1-1} \qquad \text{y} \qquad (n_2 - 1)\, S_{Y_2}^2 / \sigma_{Y_2}^2 \;=\; \chi^2_{n_2-1}$$

Y de acuerdo con la propiedad aditiva de χ^2 (ver Apéndice 5, ecuación [5.17]): $\chi^2_{n_1-1} + \chi^2_{n_2-1} = \chi^2_{n_1+n_2-2}$. Por tanto, si asumimos $\sigma^2_{Y_1} = \sigma^2_{Y_2} = \sigma^2_{Y}$, tendremos

$$\chi^2_{n_1+n_2-1} \;=\; (n_1 - 1)\, S_{Y_1}^2 / \sigma_{Y_1}^2 \;+\; (n_2 - 1)\, S_{Y_2}^2 / \sigma_{Y_2}^2 \qquad [11.12]$$

Recordemos ahora (ver Apéndice 9, ecuación [9.15]) que siendo Z una variable normal $N(0,\, 1)$ y X^2 una variable *ji*-cuadrado con *gl* grados de libertad, la transformación

$$T \;=\; \frac{Z}{\sqrt{X^2/(gl)}}$$

sigue el modelo de distribución de probabilidad *t* de *Student* con *gl* grados de libertad. En consecuencia, de acuerdo con [11.2], [11.12] y [9.15], la transformación

$$T \;=\; \frac{\dfrac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_{Y_1} - \mu_{Y_2})}{\sqrt{\sigma^2_{Y_1}/n_1 + \sigma^2_{Y_2}/n_2}}}{\sqrt{\dfrac{(n_1-1)\, S^2_{Y_1}/\sigma^2_{Y_1} \;+\; (n_2-1)\, S^2_{Y_2}/\sigma^2_{Y_2}}{n_1 + n_2 - 2}}} \qquad [11.13]$$

se distribuirá según el modelo *t* de *Student* con $n_1 + n_2 - 2$ grados de libertad. Simplificando [11.13] asumiendo $\sigma^2_{Y_1} = \sigma^2_{Y_2} = \sigma^2_{Y}$, se obtiene

$$T \;=\; \frac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_{Y_1} - \mu_{Y_2})}{\sqrt{\dfrac{(n_1 - 1)\, S^2_{Y_1} + (n_2 - 1)\, S^2_{Y_2}}{n_1 + n_2 - 2}\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$$

que no es otra cosa que el estadístico *T* de Student ya propuesto en [11.7] para el contraste sobre dos medias independientes asumiendo varianzas iguales.

El paso de [11.14] a [11.3] sólo es posible si las varianzas poblacionales $\sigma^2_{Y_1}$ y $\sigma^2_{Y_2}$ son iguales y, consecuentemente, $\sigma^2_{Y_1}/\sigma^2_{Y_2} = 1$. Sólo de esa manera [11.14] tiene solución y por esta razón decimos que, al aplicar la prueba *T* para dos medias independientes con varianzas poblacionales desconocidas, es necesario asumir que las varianzas poblacionales son iguales (supuesto éste al que nos hemos referido ya como *igualdad de varianzas* u *homocedasticidad*).

El contraste sobre igualdad de varianzas

Existen diferentes procedimientos[10] para contrastar la hipótesis de que dos varianzas poblacionales son iguales. Uno de los más simples, debido a Levene (1960), consiste en transformar los valores originales Y_{ij} (*j* se refiere al *j*-ésimo grupo: $j = 1, 2$; *i* se refiere al *i*-ésimo sujeto; por tanto: $i = 1, 2, ..., n_1$ cuando $j = 1$; $i = 1, 2, ..., n_2$ cuando $j = 2$) en puntuaciones diferenciales en valor absoluto:

$$|y_j| = |Y_{ij} - \bar{Y}_j| \qquad [11.14]$$

Se obtienen así unas nuevas puntuaciones que reflejan el grado de variabilidad presente en cada muestra (cuanto mayor sea la varianza de una muestra, mayores serán los valores absolutos de y_j y mayor, en consecuencia, su media). Sobre esas puntuaciones se aplica el estadístico *T* sobre diferencia de medias (ecuación [11.6]) para contrastar, a partir de

$$|\bar{y}_1| = \frac{1}{n_1}\sum_{i=1}^{n_1}|y_1| \quad \text{y} \quad |\bar{y}_2| = \frac{1}{n_2}\sum_{i=1}^{n_2}|y_2| \qquad [11.15]$$

la hipótesis de que ambas muestras proceden de poblaciones con la misma media, es decir,

$$H_0\!: \mu_{|y_1|} = \mu_{|y_2|} \qquad [11.16]$$

Si las varianzas poblacionales son iguales, las medias muestrales de las puntuaciones diferenciales definidas en [11.15] serán parecidas; si las varianzas poblacionales son distintas, las medias muestrales de las puntuaciones diferenciales serán distintas. Por tanto, si el estadístico *T* lleva al rechazo de la hipótesis [11.16], no podrá asumirse que las varianzas poblacionales son iguales.

Hasta ahora, en todo momento hemos hablado de la heterogeneidad de varianzas como de algo relacionado con la prueba *T* sobre diferencia de medias y, por tanto, como algo poco deseable. Pero lo cierto es que la heterogeneidad de varianzas puede constituir, ella misma, un resultado experimental relevante. Esto significa que, en ocasiones, el estudio de la variabilidad puede ser un fin en sí misma y no sólo un paso previo para la comparación de medias (ver, por ejemplo, Bryk y Raudenbush, 1988).

Imaginemos que se quiere evaluar el nivel de desarrollo cognitivo alcanzado por dos grupos de niños que han seguido programas de instrucción diferentes. Para constatar cuál de los dos grupos ha alcanzado, en promedio, mayor nivel de desarrollo, bastará con comparar las medias de ambos grupos con alguno de los procedimientos ya estudiados. Pero esta forma de proceder pasaría por alto una cuestión de cierta importancia: podría ocurrir que uno de los métodos consiguiera incrementar el nivel de desarrollo de forma generalizada (todos los niños mejoran su nivel de desarrollo) y que el otro método consiguiera el mismo objetivo con sólo unos pocos niños, aunque de forma más marcada. Estas diferencias entre ambos métodos no quedarían reflejadas en las medias, pero sí en las varianzas, por lo que sólo acompañando el contraste de medias con un contraste de varianzas podríamos obtener información acertada sobre lo que ha ocurrido.

---

[10] Ver, por ejemplo, O'Brien, 1981. En Conover, Jhonson y Jhonson (1981) se comparan 60 procedimientos diferentes para contrastar la igualdad de varianzas. La tradicional prueba *F* sobre igualdad de varianzas (Hartley, 1940, 1950) puede consultarse en Pardo y San Martín, 1998, págs. 214-215.

Ejercicios

11.1. Con el fin de estudiar la posible influencia del tipo de instrucciones sobre la ejecución de una tarea se ha seleccionado aleatoriamente una muestra de 13 sujetos. Cinco de ellos han realizado la tarea tras recibir instrucciones breves y sencillas (grupo 1); el resto, tras recibir instrucciones largas y explícitas (grupo 2). Asumiendo que las poblaciones muestreadas se distribuyen normalmente y con la misma varianza, ¿qué puede concluirse acerca del efecto del tipo de instrucciones sobre la ejecución de la tarea? (α = 0,05).

Grupos	*Sujetos*								n_j	$\bar{Y}_j$	S_{Y_j}
1	1	3	6	7	8				5	5	2,92
2	3	5	6	5	8	9	8	4	8	6	2,14

11.2. Existe evidencia empírica que apoya la hipótesis de que las mujeres que han sufrido algún tipo de abuso sexual en la infancia desarrollan en la edad adulta ciertas pautas de comportamiento que reflejan la presencia de secuelas importantes derivadas del abuso experimentado. Entre otras cosas, son más ansiosas que las mujeres que no han sufrido tal abuso y muestran con frecuencia síntomas depresivos y fóbicos. Nada se sabe, sin embargo, sobre su conducta de *afrontamiento*. Para estudiar esto último se han formado dos grupos: uno de mujeres en cuyo historial clínico existe algún episodio de abuso sexual y otro de mujeres sin la presencia de tales episodios. Tras evaluar en ambos grupos la respuesta de afrontamiento se han obtenido los siguientes resultados:

Grupo	$\bar{Y}_j$	S_Y^2	n_j
1 = Con abuso	39,5	20	20
2 = Sin abuso	43,0	15	60

¿Se puede afirmar, con α = 0,01, que las mujeres que sufren algún tipo de abuso sexual (grupo experimental) puntúan en *afrontamiento* más bajo que las mujeres que no han sufrido tal abuso?

11.3. Un investigador cree que los introvertidos y los extrovertidos se diferencian en la resistencia de unos y otros a experimentar el síndrome de indefensión aprendida (déficit cognitivo, motivacional y afectivo) que aparece en ocasiones tras una experiencia aversiva inescapable. Para comprobarlo, ha presentado una tarea de resolución de problemas a una muestra de 22 introvertidos y a otra de 16 extrovertidos. La peculiaridad de estos problemas es que no tienen solución. Tras esto, todos los sujetos pasan por una nueva situación en la que se les presenta un conjunto de problemas parecidos a los anteriores pero con la diferencia de que éstos sí tienen solución. La expectativa del investigador es que los sujetos que hayan creado indefensión en la primera tarea rendirán, en la segunda, peor que los sujetos que no la hayan creado. El investigador anota el número de problemas resueltos por cada sujeto en la segunda tarea y obtiene una media de 3,5 y una desviación típica de 1,8 en el grupo de introvertidos, y una

media de 6,3 y una desviación típica de 3,2 en el grupo de extrovertidos. Con estos resultados y utilizando $\alpha = 0{,}05$, ¿puede concluirse que los introvertidos y los extrovertidos difieren en su resistencia a experimentar indefensión?

11.4. Un educador cree que el orden en el que se presentan las preguntas de un examen afecta al rendimiento de los sujetos. Para obtener alguna evidencia sobre su sospecha, ha seleccionado un conjunto de preguntas y las ha ordenado por su nivel de dificultad estimado. Basándose en esta ordenación, ha preparado dos exámenes: el primero (A) con las preguntas ordenadas en dificultad creciente y el segundo (B) con las preguntas ordenadas en dificultad decreciente. Ha dividido aleatoriamente una clase de 20 sujetos en dos grupos de 10 y a uno de ellos le ha pasado el examen A y al otro el B. Los resultados obtenidos se ofrecen en la siguiente tabla. ¿Puede concluirse, con $\alpha = 0{,}05$, que el orden de las preguntas afecta al rendimiento en el examen?

Examen	*Sujetos*										*Suma*
A	82	78	83	95	90	65	90	75	72	70	800
B	72	68	78	66	75	55	60	56	80	70	680

11.5. El los ejercicios 11.1 al 11.4 se han aplicado contrastes de hipótesis basados en el estadístico T de Student para muestras independientes. La distribución muestral del estadístico T se aproxima a la distribución t cuando las muestras proceden de poblaciones normales. En los cuatro ejercicios hemos asumido la normalidad de las poblaciones. ¿Es correcto asumir normalidad en los cuatro casos?

11.6. ¿Entre qué límites puede estimarse que se encuentra la diferencia en el afrontamiento medio de los dos grupos del ejercicio 11.2? ¿Qué puede hacerse para mejorar la precisión del intervalo?

11.7. Al contrastar H_0: $\mu_1 - \mu_2 \leq 0$ frente a H_1: $\mu_1 - \mu_2 > 0$ se ha obtenido, para el estadístico del contraste, un valor $T = -2$. Sabiendo que $P(T \leq -2) = 0{,}045$ y utilizando $\alpha = 0{,}05$:

a. ¿Qué decisión debe tomarse sobre H_0?
b. ¿Por qué? ( ) $-2 < 0{,}05$ ( ) $0{,}045 < 0{,}05$ ( ) $0{,}955 > 0{,}05$ ( ) $0{,}955 > 0{,}95$
c. ¿A partir de qué nivel de significación podría rechazarse H_0?

11.8. Al contrastar H_0: $\mu_1 - \mu_2 \geq 4$ frente a H_1: $\mu_1 - \mu_2 < 4$ se ha obtenido para el estadístico del contraste, un valor $T = -2$. Sabiendo que $P(T \leq -2) = 0{,}025$ y utilizando $\alpha = 0{,}05$:

a. ¿Qué decisión debe tomarse sobre H_0?
b. ¿Por qué?: ( ) $-2 < 4$ ( ) $0{,}025 < 0{,}05$ ( ) $0{,}05 < 0{,}975$ ( ) $0{,}975 > 0{,}025$
c. ¿A partir de qué nivel de significación podría rechazarse H_0?

11.9. Un investigador afirma que, entre los jóvenes de 15 a 20 años, ellas fuman más que ellos. Tras efectuar una encuesta a una muestra aleatoria de 100 jóvenes y registrar el número de cigarrillos/día, ha puesto a prueba la hipótesis nula H_0: $\mu_{\text{mujeres}} \leq \mu_{\text{hombres}}$ frente a la alternativa H_1: $\mu_{\text{mujeres}} > \mu_{\text{hombres}}$ y ha obtenido una diferencia tipificada $T = 1{,}984$. La siguiente tabla recoge algunos valores de la distribución muestral de T:

T	0,526	0,845	1,290	1,660	1,984	2,365	2,626
$F(T)\,|\,H_0$	0,700	0,800	0,900	0,950	0,975	0,990	0,995

En este escenario, y con $\alpha = 0{,}01$, ¿cuál de las siguientes decisiones es correcta?

a. Rechazar H_0 porque 1,984 > 1,660.
b. Mantener H_0 porque 1,984 > 1,660.
c. Rechazar H_0 porque $P(T > 1{,}984) < 0{,}01$.
d. Mantener H_0 porque $P(T > 1{,}984) < 0{,}01$.
e. Rechazar H_0 porque 1,984 cae en zona crítica.

Soluciones

11.1. Contraste sobre dos medias independientes asumiendo varianzas iguales.

1. *Hipótesis*: $H_0: \mu_1 = \mu_2$
   $H_1: \mu_1 \neq \mu_2$ (contraste bilateral)
2. *Supuestos*: se tienen dos muestras aleatorias extraídas de poblaciones normales cuyas varianzas se asumen iguales.
3. *Estadístico del contraste* (ecuación [11.7]):

$$T = \frac{(5-6)-0}{\sqrt{[(4(2{,}92)^2 + 7(2{,}14)^2)/(5+8-2)]\,(1/5+1/8)}} = -0{,}72$$

4. *Distribución muestral*: T se distribuye según t con $n_1 + n_2 - 2 = 5 + 8 - 2 = 11$ grados de libertad.
5. *Zona crítica*: $T \leq t_{11;\,0{,}025} = -2{,}201$ y $T \geq t_{11;\,0{,}975} = 2{,}201$.
6. *Decisión*: como el valor del estadístico de contraste (−0,72) se encuentra dentro de la zona de aceptación ($-2{,}201 < -0{,}72 < 2{,}201$), se mantiene H_0. Por tanto, no es posible afirmar que el tipo de instrucciones afecte a la ejecución de la tarea.
7. *Nivel crítico*: $p = 2\,[P(T \leq -072)] \approx 2\,(0{,}25) \approx 0{,}50$.

11.2. Contraste sobre dos medias independientes no asumiendo varianzas iguales.

1. *Hipótesis*: $H_0: \mu_1 \geq \mu_2$
   $H_1: \mu_1 < \mu_2$ (contraste unilateral izquierdo)

   El contraste se plantea unilateral izquierdo porque el interés del estudio está en averiguar si el grupo 1 puntúa en afrontamiento más bajo que el grupo 2.
2. *Supuestos*: se tienen dos muestras aleatorias procedentes de poblaciones normales cuyas varianzas no se asumen iguales.
3. Estadístico del contraste (ecuación [11.10]):

$$T' = \frac{39{,}5 - 43 - 0}{\sqrt{20/20 + 15/60}} = -3{,}13$$

4. Distribución muestral: T' se distribuye según t con 29 grados de libertad (t_{29}), pues aplicando la ecuación [11.11] se obtiene

$$gl' = \frac{(20/20 + 15/60)^2}{(20/20)^2/19 + (15/60)^2/59} = 29{,}1$$

5. Zona crítica: $T' \leq t_{29;\,0,01} = -2{,}462$.
6. Decisión: el valor del estadístico del contraste (−3,13) es menor que el punto crítico (−2,462); por tanto, se rechaza H_0. Puede concluirse que el nivel de afrontamiento medio es más bajo en las mujeres que han sufrido abuso sexual (grupo 1) que en las mujeres que no lo han sufrido (grupo 2).
7. *Nivel crítico*: $p = P(T \leq -3{,}13) < 0{,}005$.

11.3. Contraste sobre dos medias independientes no asumiendo varianzas iguales.

1. *Hipótesis*: $H_0: \mu_{\text{introvertidos}} = \mu_{\text{extrovertidos}}$
   $H_1: \mu_{\text{introvertidos}} \neq \mu_{\text{extrovertidos}}$ (contraste bilateral)
2. *Supuestos*: se tienen dos muestras aleatorias procedentes de poblaciones normales cuyas varianzas no se asumen iguales.
3. Estadístico del contraste (ecuación [11.10]):

$$T' = \frac{3{,}5 - 6{,}3 - 0}{\sqrt{1{,}8^2/22 + 3{,}2^2/16}} = -3{,}16$$

4. Distribución muestral: T' se distribuye según t con 22 grados de libertad (t_{22}), pues aplicando la ecuación [11.11] se obtiene

$$gl' = \frac{(1{,}8^2/22 + 3{,}2^2/16)^2}{(1{,}8^2/22)^2/21 + (3{,}2^2/16)^2/15} = 21{,}9$$

5. Zona crítica: $T \leq t_{22;\,0,025} = -2{,}074$ y $T \geq t_{22;\,0,975} = 2{,}074$.
6. Decisión: como el valor del estadístico del contraste (−3,16) es menor que el punto crítico de la cola izquierda (−2,074), se rechaza H_0. Por tanto, puede concluirse que el rendimiento medio de los introvertidos y los extrovertidos en la tarea de solución de problemas no es el mismo, lo cual significa que difieren en su resistencia a experimentar indefensión.
7. *Nivel crítico*: $p = 2\,[P(T \leq -3{,}16)] < 2\,(0{,}005) \rightarrow p < 0{,}01$.

11.4. Contraste sobre dos medias independientes no asumiendo varianzas iguales.

1. *Hipótesis*: $H_0: \mu_A = \mu_B$
   $H_1: \mu_A \neq \mu_B$ (contraste bilateral)

   El contraste se plantea unilateral izquierdo porque el interés del estudio está en averiguar si el grupo experimental puntúa en afrontamiento más bajo que el grupo control.
2. *Supuestos*: se tienen dos muestras aleatorias procedentes de poblaciones normales cuyas varianzas no se asumen iguales.
3. Estadístico del contraste (ecuación [11.10]):

$$\bar{A} = 800/10 = 80$$

$$\bar{B} = 680/9 = 68$$

$$S_A^2 = \frac{(82-80)^2 + (78-80)^2 + \cdots + (70-80)^2}{9} = 95{,}11$$

$$S_A^2 = \frac{(72-68)^2 + (68-68)^2 + \cdots + (70-68)^2}{9} = 77{,}11$$

$$T' = \frac{80 - 68 - 0}{\sqrt{95{,}11/10 + 77{,}11/10}} = 2{,}89$$

4. Distribución muestral: T' se distribuye según t con 18 grados de libertad (t_{18}), pues aplicando la ecuación [11.11] se obtiene

$$gl' = \frac{(95{,}11/10 + 77{,}11/10)^2}{(95{,}11/10)^2/9 + (77{,}11/10)^2/9} = 17{,}81$$

5. Zona crítica: $T \leq t_{18;\,0{,}025} = -2{,}101$ y $T \geq t_{22;\,0{,}975} = 2{,}101$.
6. Decisión: como el valor del estadístico del contraste (2,89) es mayor que el punto crítico de la cola derecha (2,101), se rechaza H_0. Por tanto, puede concluirse que el rendimiento medio en los dos exámenes no es el mismo. Este resultado apoya la sospecha del educador de que el orden de las preguntas afecta al rendimiento de los sujetos.
7. *Nivel crítico*: $p = 2\,[P\,(T \geq 2{,}89)] \approx 2\,(0{,}05) \approx 0{,}01$.

11.5. En el ejercicio 11.1 se dice explícitamente que las poblaciones muestreadas son normales; por tanto, es correcto asumir normalidad. En los ejercicios 11.2 y 11.3 no se conoce la forma de las poblaciones muestreadas, pero en ambos casos los tamaños muestrales son lo bastante grandes como para poder asumir normalidad sin que esto perjudique el comportamiento del estadístico T (ver el apartado *Independencia, normalidad e igualdad de varianzas*). Por último, aunque los tamaños muestrales utilizados en el ejemplo del ejercicio 11.4 no son grandes, ocurre que, además de ser iguales, las distribuciones son razonablemente simétricas y no hay presencia de casos atípicos (ambas cosas pueden comprobarse fácilmente con unos diagramas de caja); por tanto, tampoco en este caso es incorrecto asumir que las poblaciones muestreadas son normales.

11.6. Intervalo de confianza para la diferencia entre dos medias independientes (con $\alpha = 0{,}01$). Puesto que no se está asumiendo que las varianzas poblacionales sean iguales, el error típico de la diferencia entre las medias ($\hat{\sigma}_{\bar{Y}_1 - \bar{Y}_2}$) hay que calcularlo a partir de la ecuación [11.9] (no a partir de la ecuación [11.4]). En realidad, este error típico ya lo hemos calculado (es el denominador del estadístico T' del ejercicio 11.2):

$$\hat{\sigma}_{\bar{Y}_1 - \bar{Y}_2} = \sqrt{20/20 + 15/60} = 1{,}12$$

Por tanto,

$$IC_{|\mu_{Y_1} - \mu_{Y_2}|} = |\bar{Y}_1 - \bar{Y}_2| \pm t_{gl;\,1-\alpha/2}\,\hat{\sigma}_{\bar{Y}_1 - \bar{Y}_2} = |39{,}5 - 43| \pm t_{29;\,0{,}995}\,(1{,}12) =$$
$$= 3{,}5 \pm 2{,}756\,(1{,}12) = 3{,}5 \pm 3{,}09 = (0{,}41;\ 6{,}59)$$

La estrategia idónea para aumentar la precisión de un intervalo consiste en aumentar el tamaño de las muestras. Pero, como eso es algo que no podemos hacer aquí, nos queda la alternativa de disminuir el nivel de confianza. Si en lugar de utilizar un nivel de confianza de 0,99 se utiliza un nivel de confianza de 0,95, el intervalo se estrecha hasta (1,21; 5,79).

11.7. *a*. Mantenerla.
b. Porque $0{,}955 > 0{,}05$ (debe repararse en el hecho de que el estadístico del contraste y la zona crítica se encuentran en colas opuestas de la distribución muestral).
c. Nivel crítico: $p = 0{,}955$.

11.8. *a*. Rechazarla.
b. Porque $0{,}025 < 0{,}05$.
c. Nivel crítico: $p = 0{,}025$.

11.9. Con $\alpha = 0{,}01$, el valor del estadístico del contraste (1,984) cae fuera de la zona crítica porque, siendo el contraste unilateral derecho, el punto crítico vale 2,365. La decisión que debe tomarse es, por tanto, mantener H_0. Esto descarta como correctas las alternativas *a*, *c* y *e*. En la alternativa *b* se está tomando

la decisión correcta pero basada en un argumento equivocado: aunque es cierto que 1,984 > 1,660, esto no tiene nada que ver con el motivo por la cual se mantiene H_0. Y en la alternativa *d* ocurre algo parecido: se está tomando la decisión correcta basada en un argumento equivocado, pues cuando un estadístico de contraste cae en la zona crítica no debe mantenerse H_0, sino rechazarse. En conclusión, ninguna de las alternativas es correcta.

12

Inferencia con dos variables cuantitativas

Muestras relacionadas

Dos variables cuantitativas pueden obtenerse de diferentes maneras. La más habitual consiste en tomar dos medidas a los mismos sujetos, bien midiendo dos variables distintas (altura y peso, calificaciones en lengua y en matemáticas, etc.), bien midiendo la misma variable en dos momentos distintos (el nivel de ansiedad antes y después de un examen, el peso antes y después de participar en un programa de adelgazamiento, etc.).

También se tienen dos variables cuantitativas cuando, en lugar de utilizar los mismos sujetos, se utilizan *pares* de sujetos que comparten alguna característica que pueda resultar relevante para el análisis. Por ejemplo, en un estudio sobre satisfacción conyugal se puede medir el grado de satisfacción en los dos miembros de cada pareja; en un estudio acerca del efecto de la herencia y el ambiente en la inteligencia se puede medir el cociente intelectual a pares de gemelos; etc. En este tipo de estudios, los miembros de cada par están asociados mediante algún criterio considerado relevante para el análisis; y, aunque cada miembro del par contribuye con una puntuación, cada par constituye una unidad de análisis.

Tanto si se utilizan los mismos sujetos como si se utilizan sujetos emparejados, lo que caracteriza a este tipo de datos es que no son independientes entre sí; y no lo son porque, tanto en el caso de dos puntuaciones pertenecientes al mismo sujeto como en el de puntuaciones pertenecientes a dos sujetos emparejados, el conocimiento de una de las puntuaciones del par permite saber algo de la otra puntuación del mismo par: los buenos estudiantes tienden a obtener puntuaciones altas tanto en lengua como en matemáticas; los sujetos muy ansiosos tienden a experimentar un nivel alto en ansiedad tanto antes como después de un examen; el nivel de satisfacción conyugal de los dos miembros de la misma pareja cabe esperar que sea parecido; el cociente intelectual de un sujeto cabe esperar que se parezca más al de su gemelo que al de otro sujeto cualquiera; etc. Puede que una puntuación no diga mucho de la otra, pero es seguro que algo dice.

A los diseños que permiten recoger este tipo de información (dos puntuaciones a los mismos sujetos o a dos sujetos emparejados; y lo mismo vale decir de tres o más puntuaciones,

aunque aquí nos estemos limitando a dos) se les llama *diseños con los mismos sujetos* o *diseños intrasujeto* (en el caso de sujetos emparejados –o tríos, o cuartetos, etc.– también se habla de *diseños de bloques aleatorios con un sujeto por nivel y bloque*). En el contexto del análisis de datos se habla, queriendo significar exactamente lo mismo, de *medidas repetidas* o *muestras relacionadas*.

Comparar o relacionar

Al trabajar con dos variables cuantitativas el interés del análisis puede orientarse hacia dos objetivos bien diferentes: *compararlas* o *relacionarlas*.

Ya hemos señalado que, cuando hay variables categóricas por medio, comparar y relacionar son la misma cosa: si hay diferencias, hay relación; y a la inversa. Por ejemplo, si la proporción de fumadores difiere de la proporción de fumadoras, entonces la variable *sexo* está relacionada con la variable *tabaquismo*; si el salario medio de las personas con alto nivel educativo difiere del salario medio de las personas con bajo nivel educativo, entonces la variable *salario* está relacionada con la variable *nivel educativo*.

Pero cuando las dos variables son cuantitativas, comparar y relacionar significan cosas muy distintas. La comparación se basa en los centros (promedios) de las variables; la relación se basa en la forma de variar las puntuaciones. Pensemos en las valoraciones que hacen dos entrevistadores de los aspirantes a un puesto de trabajo. Comparar las valoraciones significa averiguar si los dos entrevistadores tienen el listón puesto a la misma altura o, por el contrario, uno de ellos es más duro o exigente que el otro (uno de ellos puntúa sistemáticamente más bajo que el otro). Relacionar las valoraciones significa averiguar si los entrevistadores, independientemente de dónde tengan colocado el listón, coinciden en las valoraciones que hacen, es decir, coinciden en quién es el mejor candidato, el segundo mejor, etc.

Una importante consecuencia de esta diferencia entre comparar y relacionar es que sólo tiene sentido comparar variables que se encuentran en la misma métrica. Las calificaciones obtenidas en lengua pueden compararse con las obtenidas en matemáticas (se están comparando puntuaciones que se encuentran en una métrica que va de 0 a 10 puntos); el peso de los sujetos antes de iniciar un programa de adelgazamiento puede compararse con el peso de esos mismos sujetos después de acabado el programa (se están comparando kg). Pero no tiene ningún sentido comparar el nivel educativo (medido en años de formación académica) con el salario anual (medido en euros); como tampoco tiene sentido comparar la altura de los sujetos (medida en cm) con su peso (medido en kg).

Por el contrario, siempre es posible relacionar dos variables cuantitativas independientemente de la métrica en la que se encuentren. Relacionar la altura medida en cm con el peso medido en kg significa averiguar si los sujetos que más puntúan en altura son también los que más puntúan en peso; relacionar el nivel educativo medido en años de formación académica con el salario anual medido en euros significa averiguar si los sujetos con mayor nivel educativo son también los que tienen mayor salario.

En este capítulo se presta atención a ambas cosas: cómo comparar dos variables cuantitativas y cómo relacionarlas. Para compararlas estudiaremos la *prueba T de Student para muestras relacionadas* o dependientes; para relacionarlas, el *coeficiente de correlación de Pearson*. Estos dos estadísticos son, quizá, los más utilizados al analizar dos variables cuantitativas. Sin embargo, no son los únicos. En el segundo volumen tendremos ocasión de estudiar otros estadísticos.

La prueba *T* de Student para muestras relacionadas

La prueba *T* para muestras relacionadas sirve para contrastar la hipótesis de igualdad de medias cuando las poblacionales con las que se trabaja no son independientes. No debemos perder de vista que estamos trabajando con dos variables cuantitativas y que el objetivo del análisis es compararlas.

En el contraste sobre dos medias independientes estudiado en el capítulo anterior, se asume que $\bar{Y}_1$ e $\bar{Y}_2$ son variables aleatorias independientes. En la práctica, esta independencia queda garantizada calculando cada media en una muestra seleccionada aleatoria e independientemente de su respectiva población.

Ahora también tenemos dos poblaciones (Y_1 e Y_2) y una muestra aleatoriamente seleccionada de cada población. Pero en cada extracción, las dos observaciones seleccionadas no se consideran independientes porque corresponden al mismo sujeto o a dos sujetos emparejados mediante algún vínculo relevante para el análisis. En este nuevo escenario, puesto que las puntuaciones de cada par están relacionadas, pueden transformarse en diferencias:

$$D = Y_1 - Y_2 \qquad [12.1]$$

De esta forma, a cada sujeto o par le corresponde una única puntuación. Estas puntuaciones (diferencias) valen cero cuando las dos puntuaciones del mismo par son iguales, son negativas cuando la primera puntuación del par es menor que la segunda y son positivas cuando la primera puntuación del par es mayor que la segunda. En el caso de diseños antes-después o pre-post, la diferencia entre las puntuaciones de cada par refleja el cambio (pérdida o ganancia) entre los dos momentos.

Así pues, al estudiar dos medias relacionadas puede considerarse que tenemos una única población (la población de diferencias D) con media μ_D y varianza σ_D^2. Seleccionando de esa población una muestra aleatoria de n observaciones y definiendo el estadístico

$$\bar{D} = \sum_i^n D_i / n \qquad \text{(o, lo que es lo mismo, } \bar{D} = \bar{Y}_1 - \bar{Y}_2\text{)} \qquad [12.2]$$

obtenemos una variable con valor esperado y varianza (ver Apéndice 6):

$$E(\bar{D}) = \mu_D$$
$$V(\bar{D}) = \sigma_{\bar{D}}^2 = \sigma_D^2 / n \qquad \rightarrow \qquad \sigma_{\bar{D}} = \sigma_D / \sqrt{n} \qquad [12.3]$$

Por tanto, nos encontramos en una situación idéntica a la descrita en el Capítulo 9 a propósito del contraste sobre una media (prueba *T* para una muestra). Si la población de las diferencias D es normal, entonces la distribución muestral de $\bar{D}$ también es normal con los parámetros definidos en [12.3]. Consecuentemente,

$$Z = \frac{\bar{D} - E(\bar{D})}{\sigma_{\bar{D}}} = \frac{\bar{D} - \mu_D}{\sigma_D / \sqrt{n}} \sim N(0,1) \qquad [12.4]$$

Lo cual significa que podemos utilizar la transformación Z y la distribución normal tipificada para conocer las probabilidades asociadas a los diferentes valores del estadístico $\bar{D}$.

Ahora bien, como la varianza de la población de diferencias es, por lo general, un valor desconocido, es necesario estimarlo mediante

$$\hat{\sigma}^2_{\bar{D}} = S^2_D / n, \qquad \text{con} \quad S^2_D = \sum_i (D_i - \bar{D})^2 / (n-1) \qquad [12.5]$$

En cuyo caso, la transformación

$$T = \frac{\bar{D} - \mu_D}{S_D / \sqrt{n}} \qquad [12.6]$$

ya no se ajusta a la distribución normal tipificada, sino a la distribución t de Student con $n-1$ grados de libertad. Tenemos, por tanto, todo lo necesario para diseñar un contraste de hipótesis sobre la diferencia entre dos medias relacionadas. El Cuadro 12.1 resume los pasos de este contraste.

Cuadro 12.1. Resumen del contraste sobre dos medias relacionadas (prueba *T* de *Student* para muestras relacionadas o dependientes)

1. *Hipótesis*[1]:
   a. Contraste bilateral: H_0: $\mu_D = k_0$; H_1: $\mu_D \neq k_0$.
   b. Contraste unilateral derecho: H_0: $\mu_D \leq k_0$; H_1: $\mu_D > k_0$.
   c. Contraste unilateral izquierdo: H_0: $\mu_D \geq k_0$; H_1: $\mu_D < k_0$.
2. *Supuestos*: muestra aleatoria de n diferencias procedentes de una población normal (el supuesto de *normalidad* va perdiendo importancia conforme el tamaño muestral va aumentando)[2].
3. *Estadístico del contraste* (ecuación [12.6]):

$$T = \frac{\bar{D} - \mu_D}{S_D / \sqrt{n}}$$

4. *Distribución muestral*: T se distribuye según t con $n-1$ grados de libertad (t_{n-1}).
5. *Zona crítica*:
   a. Contraste bilateral: $T \leq t_{n-1;\, \alpha/2}$ y $T \geq t_{n-1;\, 1-\alpha/2}$.
   b. Contraste unilateral derecho: $T \geq t_{n-1;\, 1-\alpha}$.
   c. Contraste unilateral izquierdo: $T \leq t_{n-1;\, \alpha}$.
6. *Regla de decisión*: se rechaza H_0 si el estadístico del contraste cae en la zona crítica; en caso contrario, se mantiene.

[1] Al igual que en el contraste sobre dos medias independientes, generalmente $k_0 = 0$, pues la hipótesis nula que habitualmente interesará contrastar será H_0: $\mu_D = 0$. Además, debe tenerse en cuenta que $\mu_D = \mu_{Y_1} - \mu_{Y_2}$. Por tanto, la hipótesis nula que habitualmente interesará contrastar podrá formularse como H_0: $\mu_{Y_1} = \mu_{Y_2}$.

[2] En lo relativo a cómo de grande tiene que ser el tamaño muestral para que el supuesto de normalidad deje de ser un problema, sirve aquí lo ya dicho en el apartado *Independencia y normalidad* del Capítulo 9.

7. *Nivel crítico* (valor p):
   a. Contraste bilateral: $p = 2[P(T \geq |T_h|)]$, siendo T_h el valor muestral concreto que toma el estadístico T.
   b. Contraste unilateral derecho: $p = P(T \geq T_h)$.
   c. Contraste unilateral izquierdo: $p = P(T \leq T_h)$.
8. *Intervalo de confianza*: $IC_{\mu_D} = \bar{D} \pm t_{n-1;\, 1-\alpha/2}\, S_D/\sqrt{n}$ [12.7]

Ejemplo. La prueba T de Student para muestras relacionadas

En un estudio diseñado para probar el efecto de un tratamiento antidepresivo mixto (fluoxetina + psicoterapia), se ha utilizado una muestra aleatoria de 14 pacientes con depresión. A todos ellos se les ha aplicado la escala de depresión de Hamilton (Y) en dos momentos: justo antes de iniciar el tratamiento (línea base o pre-test) y tras 12 semanas de tratamiento (post-test). La Tabla 12.1 muestra los resultados que han obtenido los 14 sujetos en las dos mediciones llevadas a cabo. El objetivo del estudio es averiguar si las puntuaciones en la escala disminuyen tras aplicar el tratamiento ($\alpha = 0{,}05$).

Tabla 12.1. Puntuaciones en la escala de depresión de Hamilton

Sujetos	1	2	3	4	5	6	7	8	9	10	11	12	13	14	$\bar{Y}_j$
Y_1 = Pre-test	24	38	21	14	19	31	34	33	22	16	17	20	18	23	23,57
Y_2 = Post-test	15	22	21	17	11	6	15	20	8	9	5	19	7	8	13,07

Tenemos dos conjuntos de puntuaciones que se han obtenido al medir dos veces en los mismos sujetos (muestras relacionadas) una variable cuantitativa (Y = *puntuaciones en la escala Hamilton*). Tenemos, por tanto, dos variables cuantitativas[3]. Para averiguar si las puntuaciones del post-test han disminuido respecto de las del pre-test, vamos a comparar sus promedios con la prueba T para muestras relacionadas:

1. *Hipótesis*: H_0: $\mu_{\text{antes}} \leq \mu_{\text{después}}$; H_1: $\mu_{\text{antes}} > \mu_{\text{después}}$ (contraste unilateral derecho).
2. *Supuestos*: asumimos que la muestra de 14 diferencias se han seleccionado aleatoriamente de una población normal.
3. *Estadístico del contraste*. Los cálculos realizados en la Tabla 12.2 permiten obtener fácilmente los estadísticos que necesitamos:

$$\bar{D} = \sum_i^n D_i/n = 147/14 = 10{,}5 \quad (\text{también, } \bar{D} = \bar{Y}_1 - \bar{Y}_2 = 23{,}57 - 13{,}07 = 10{,}5)$$

$$S_D^2 = \sum_i (D_i - \bar{D})^2/(n-1) = 757{,}5/13 = 58{,}27 \quad \rightarrow \quad S_D = \sqrt{58{,}27} = 7{,}63$$

[3] Debe tenerse en cuenta que se tienen dos variables cuantitativas tanto si se miden dos variables distintas como si se mide la misma variable dos veces (ver, al comienzo de este mismo capítulo, el apartado *Muestras relacionadas*).

Tabla 12.2. Cálculos realizados con los datos de la Tabla 12.1

Sujetos	1	2	3	4	5	···	13	14	
Y_1 = Pre-test	24	38	21	14	19	···	18	23	
Y_2 = Post-test	15	22	21	17	11	···	7	8	*Total*
D	9	16	0	−3	8	···	11	15	147
$(D-\bar{D})^2$	2,25	30,25	110,25	182,25	6,25	···	0,25	20,25	757,5

$$T = \frac{\bar{D}-\mu_D}{S_D/\sqrt{n}} = \frac{10{,}5-0}{7{,}63/\sqrt{14}} = \frac{10{,}5}{2{,}04} = 5{,}15.$$

4. *Distribución muestral*: T se distribuye según t con $n-1 = 14-1 = 13$ grados de libertad.
5. *Zona crítica*: $T \geq t_{13;\,0,95} = 1{,}771$.
6. *Decisión*: como $5{,}15 > 1{,}771$, se rechaza H_0. Por tanto, puede concluirse que la media del post-test es menor que la del pre-test.
7. *Nivel crítico*: $p = P(T \geq 5{,}15) < 0{,}001$.
8. *Intervalo de confianza* (con $S_{D_Y}/\sqrt{n} = 2{,}04$; denominador del estadístico T):

$$IC_{\mu_D} = \bar{D} \pm t_{n-1;\,1-\alpha/2}\, S_D/\sqrt{n} = 10{,}5 \pm t_{13;\,0,975}(2{,}04) =$$
$$= 10{,}5 \pm 2{,}16\,(2{,}04) = 10{,}5 \pm 4{,}41 = (6{,}09;\ 14{,}91).$$

Estimamos, con una confianza del 95%, que, en la población, la diferencia entre la media del pre-test y la del post-test se encuentra aproximadamente entre 6 y 15 puntos.

La prueba *T* de Student para muestras relacionadas con SPSS

La prueba T de Student para muestras relacionadas está disponible en la opción **Comparar medias > Prueba *T* para muestras relacionadas** del menú **Analizar**. La lista de variables del cuadro de diálogo principal únicamente muestra las variables con formato numérico.

Para llevar a cabo un contraste con las especificaciones que el procedimiento tiene establecidas por defecto: trasladar las dos variables cuyas medias se desea comparar a la lista **Variables relacionadas**. Las variables deben trasladarse a esta lista *por pares*; es decir, es necesario marcar dos variables de la lista de variables para que el botón flecha esté activo. El procedimiento genera una prueba T por cada par seleccionado.

Las opciones del procedimiento permiten controlar el nivel de confianza con el que se desea trabajar y el tratamiento que se desea dar a los valores perdidos. La opción **Intervalo de confianza: *k* %** permite establecer, en escala porcentual, el *nivel de confianza* $(1-\alpha)$ con el que se desea obtener el intervalo de confianza para la diferencia entre las medias. El valor de k es, por defecto, 95, pero puede introducirse cualquier otro valor comprendido entre 0,01 y 99,99. En las opciones del recuadro **Valores perdidos** se puede elegir entre dos formas distintas de tratar los casos con valores perdidos: (1) la opción **Excluir casos según análisis** excluye de cada contraste (de cada prueba T) los casos con valor perdido en cualquiera de las dos variables

que están siendo contrastadas; (2) La opción **Excluir casos según lista** excluye de todos los contrastes (de todas las pruebas *T* solicitadas) los casos con algún valor perdido en cualquiera de las variables seleccionadas en la lista **Variables relacionadas**.

Ejemplo. La prueba T de Student para muestras relacionadas con SPSS

Este ejemplo muestra cómo contrastar hipótesis sobre dos medias utilizando el procedimiento **Prueba *T* para muestras relacionadas** con los datos ya analizados en el ejemplo anterior (ver Tabla 12.1). Para ello, hemos reproducido los datos de la tabla creando dos variables a las que hemos llamado *antes* y *después*, con etiquetas 1 = «Pre-test» y 2 = «Post-test» (los datos se encuentran en el archivo *Tabla 12.1 hamilton*, en la página web del manual). Para contrastar la hipótesis de igualdad de medias:

➤ Seleccionar la opción **Comparar medias > Prueba *T* para muestras relacionadas** del menú **Analizar** y trasladar las variables *antes* y *después* a la lista **Variables relacionadas**.

Aceptando estas selecciones, el *Visor* ofrece los resultados que muestran las Tablas 12.3 y 12.4. La Tabla 12.3 informa, para cada variable, de la media, del numero de casos válidos, de la desviación típica insesgada y del error típico de la media (la desviación típica dividida por la raíz cuadrada del número de casos).

Tabla 12.3. Estadísticos descriptivos del procedimiento *Prueba T para muestras relacionadas*

	Media	N	Desviación típ.	Error típ. de la media
Pre-test	23,57	14	7,480	1,999
Post-test	13,07	14	6,032	1,612

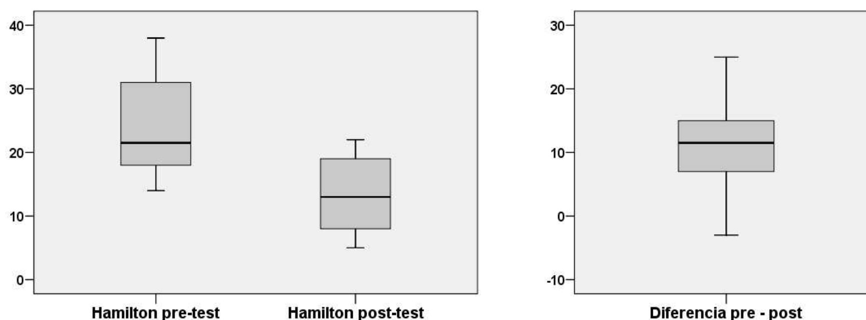
La Tabla 12.4 comienza ofreciendo tres estadísticos referidos a las *diferencias* entre cada par de puntuaciones: la media, la desviación típica y el error típico de la media. La siguiente columna contiene el intervalo de confianza para la diferencia entre las medias: puede estimarse, con una confianza del 95%, que la diferencia media entre el pre-test y el post-test se encuentra entre 6,09 y 14,91 puntos. Las tres últimas columnas informan sobre el valor del estadístico *t*, sus grados de libertad (*gl*) y el nivel crítico bilateral (*sig. bilateral*; el unilateral se obtiene dividiendo el bilateral entre 2). Puesto que el valor del nivel crítico es muy pequeño (menor que 0,0005), rechazamos la hipótesis de igualdad de medias y concluimos que la media del post-test es menor que la media del pre-test.

Tabla 12.4. Resumen del procedimiento *Prueba T para muestras relacionadas*

	Diferencias relacionadas							
				95% Intervalo de confianza para la diferencia				
	Media	Desviación típica	Error típ. de la media	Inferior	Superior	t	gl	Sig. (bilateral)
Pre-test - Post-test	10,500	7,633	2,040	6,093	14,907	5,147	13	,000

La Figura 12.1 muestra los diagramas de caja correspondientes a las variables *pre-test* y *post-test* y a la diferencia entre ambas[4]. El centro (mediana) de las puntuaciones del post-test es menor que el centro de las puntuaciones del pre-test. La dispersión es parecida, aunque ligeramente mayor en el pre-test. Y en ninguna de las dos variables ni en la diferencia entre ambas se observan indicios de asimetría ni casos atípicos o extremos.

Figura 12.1. Diagramas de caja. Puntuaciones *pre* y *post* en la escala Hamilton



Relación lineal

Al principio de este capítulo hemos advertido que el análisis de dos variables cuantitativas puede orientarse hacia dos objetivos bien diferenciados: *compararlas* o *relacionarlas*. Hasta ahora nos hemos limitado a compararlas mediante la prueba *T* de Student. A partir de aquí nos centraremos en cómo relacionarlas.

Suele decirse que cuanto mayor es el nivel educativo, mayor es el nivel de renta; que los sujetos más frustrados son también más agresivos; que los niveles altos de colesterol en sangre suelen ir acompañados de dietas alimenticias ricas en grasas; que los sujetos muestran más interés por una tarea cuanto mayor es la recompensa que reciben; que el nivel de ansiedad tiene que ver con el rendimiento en una tarea; que al aumentar el consumo de alcohol aumenta la cantidad de problemas perceptivos; etc. En todos estos ejemplos se está hablando de *relación entre dos variables cuantitativas*.

En el lenguaje común, el concepto de relación tiene un significado algo vago: esto tiene que ver con aquello (correspondencia o vínculo entre dos cosas). En estadística tiene un significado más concreto. Al estudiar dos variables categóricas ya hemos hablado de relación (la hemos llamado también *asociación*) para referirnos a la diferencia entre las distribuciones condicionales: si la proporción de fumadores difiere de la proporción de fumadoras, entonces la variable *sexo* está relacionada (asociada) con la variable *tabaquismo*. Algo parecido ocurre al analizar una variable categórica y una cuantitativa: si los hombres y las mujeres difieren en altura, entonces la variable *sexo* está relacionada con la variable *altura*.

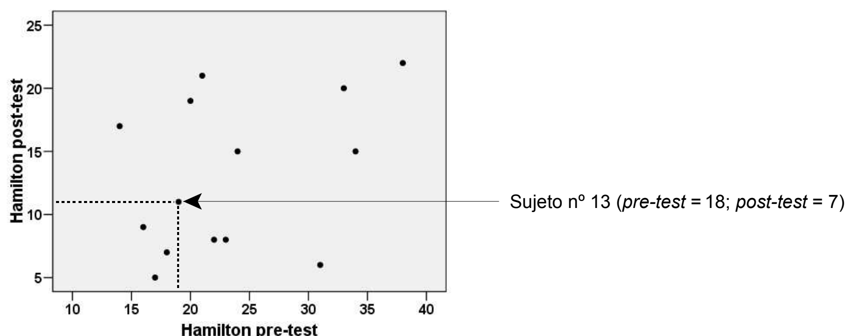
[4] Estos diagramas se han obtenido con la opción **Diagramas de caja** > **Simple (Resúmenes para distintas variables)** del menú **Gráficos**.

Al estudiar dos variables cuantitativas la situación es algo más compleja que cuando una es categórica y la otra cuantitativa. Con una variable de cada tipo sólo cabe hablar de relación (diferencias) o de no relación (no diferencias). Con dos variables cuantitativas es posible hablar de varios tipos de relación: lineal, cuadrática, etc. Para entender esto, vamos a comenzar representando gráficamente ambas variables.

Diagramas de dispersión

Un *diagrama de dispersión* es la forma más directa e intuitiva de formarse una primera impresión sobre el tipo de relación existente entre dos variables cuantitativas medidas en los mismos sujetos (o en sujetos emparejados; ver, al comienzo del capítulo, el apartado *Muestras relacionadas*). Tiene forma de una *nube de puntos* dispuesta sobre el plano definido por dos ejes cartesianos: en el eje de abscisas (horizontal) se coloca una de las variables (X), en el eje de ordenadas (vertical) se coloca la otra variable (Y), y cada par de puntuaciones (X_i, Y_i), es decir, cada sujeto, se representa con un punto. La Figura 12.2 muestra cómo se construye el diagrama de dispersión correspondiente a las variables *pre-test* y *post-test* de la Tabla 12.1. Se ha destacado en el diagrama la posición que ocupa el sujeto nº 13 de la tabla, que corresponde al punto en el que se cruza la prolongación vertical de su valor en la medida *pre-test* (18) con la prolongación horizontal de su valor en la medida *post-test* (7).

Figura 12.2. Diagrama de dispersión correspondiente a los datos de la Tabla 12.1

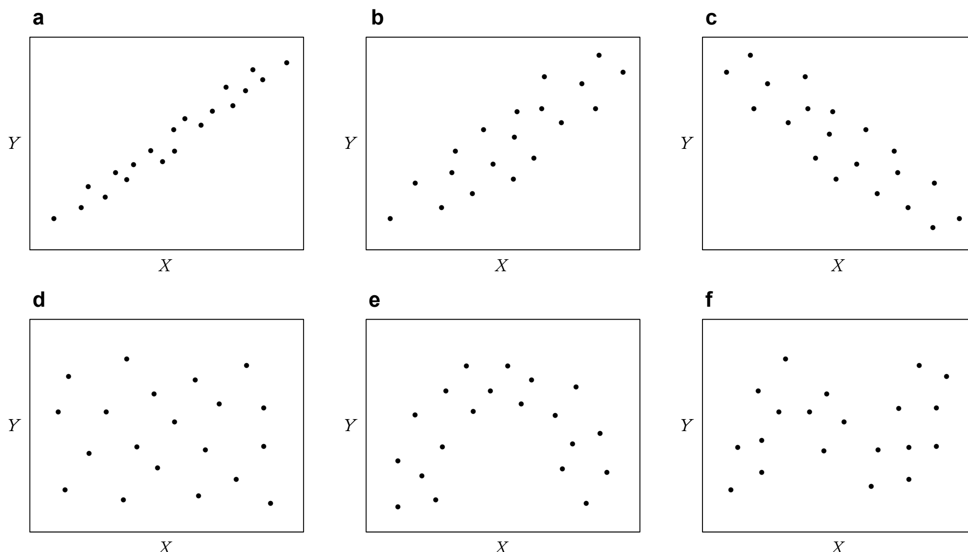


La forma de la nube de puntos informa sobre el tipo de relación existente. Para entender esto, la Figura 12.3 muestra varios diagramas de dispersión que reflejan diferentes tipos de relación. El diagrama *a* muestra una nube de puntos bastante concentrada en torno a una línea recta ascendente. Es un ejemplo típico de relación **lineal**. También hay relación lineal cuando se da la pauta de variación que muestra el diagrama *b*: los puntos siguen agrupados en torno a una línea recta ascendente, aunque de forma menos evidente que en el primer diagrama. Y también hay relación lineal cuando, como en el diagrama *b*, los puntos se agrupan en torno a una línea recta descendente.

El diagrama *d* es un claro ejemplo de ausencia de relación: los puntos están dispersos por todo el diagrama sin mostrar ninguna pauta de variación reconocible. En el diagrama *e* se observa una pauta de variación claramente no lineal: los puntos están agrupados en torno a una curva que comienza ascendiendo y termina descendiendo; a esta pauta de variación conjunta se le llama *cuadrática*. Y también en el diagrama *f* se da una pauta no lineal: los puntos están

agrupados en torno a una línea que comienza ascendiendo, a continuación desciende y termina volviendo a ascender; a esta pauta de variación se le llama *cúbica*. Aunque puede resultar interesante estudiar cualquier tipo de relación[5], la lineal es, sin duda, la más estudiada en estadística (quizá porque es la más fácil de interpretar y la que con mayor frecuencia encontramos en el mundo real) y la que va a acaparar nuestra atención aquí.

Figura 12.3. Diagramas de dispersión representando diferentes tipos de relación: (a) lineal positiva, (b) lineal positiva, (c) lineal negativa, (d) no relación o independencia, (e) curvilínea y (f) cúbica



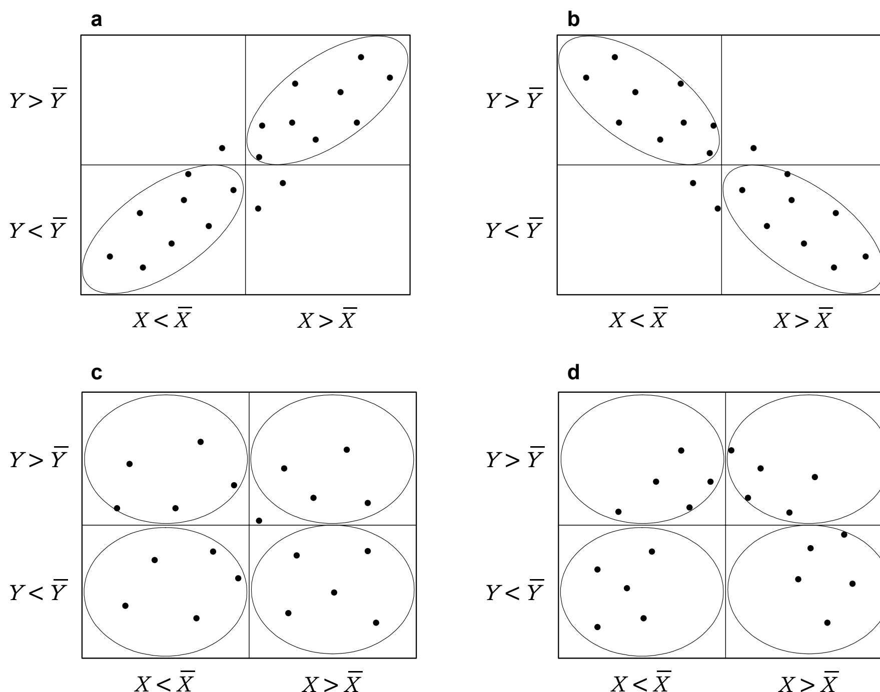
Los diagramas de dispersión de la Figura 12.4 reproducen algunos de los diagramas de la Figura 12.3, pero con información adicional: incluyen líneas que parten del centro de cada variable (para separar las puntuaciones *bajas* y *altas*) y círculos para resaltar los cuadrantes en los que se da la mayor concentración de casos.

En el diagrama *a*, las puntuaciones bajas en *X* tienden a ir acompañadas de puntuaciones bajas en *Y*, y las puntuaciones altas en *X* tienden a ir acompañadas de puntuaciones altas en *Y* (cuadrantes "altas-altas" y "bajas-bajas"); a esta pauta de variación se le llama **relación lineal positiva** o *directa* (esto es lo que ocurre, por ejemplo, con la altura y el peso, o la inteligencia y el rendimiento). En el diagrama *b*, las puntuaciones bajas en *X* tienden a ir acompañadas de puntuaciones altas en *Y*, y las puntuaciones altas en *X* tienden a ir acompañadas de puntuaciones bajas en *Y* (cuadrantes "bajas-altas" y "altas-bajas"); a esta pauta de variación se le llama **relación lineal negativa** o *inversa* (esto es lo que ocurre, por ejemplo, con la fatiga y el rendimiento, o con la velocidad de realización de una tarea y el número de errores). Cuando no existe relación lineal, bien porque no existe ningún tipo de relación, como en el diagrama *c* (altura e inteligencia, por ejemplo), bien porque la relación no es de tipo lineal,

[5] Por ejemplo, una ley muy conocida en psicología, llamada ley de Yerkes-Dodson, afirma que la relación entre ansiedad y rendimiento es cuadrática (con forma de U invertida, como la del diagrama *E* de la Figura 12.3). Esto significa que los niveles bajos y altos de ansiedad tienden a ir acompañados de niveles bajos de rendimiento, mientras que los niveles de ansiedad intermedios tienden a ir acompañados de niveles altos de rendimiento.

como en el diagrama *d* (ansiedad y rendimiento, por ejemplo), tanto las puntuaciones bajas en *X* como las altas aparecen acompañadas, indistintamente, de puntuaciones bajas y altas en *Y*. Parece claro, por tanto, que un diagrama de dispersión ayuda a formarse una idea bastante acertada sobre el *tipo de relación* existente entre dos variables cuantitativas.

Figura 12.4. Diagramas de dispersión representando diferentes tipos de relación: (a) lineal positiva, (b) lineal negativa, (c) no relación o independencia y (d) cuadrática o curvilínea



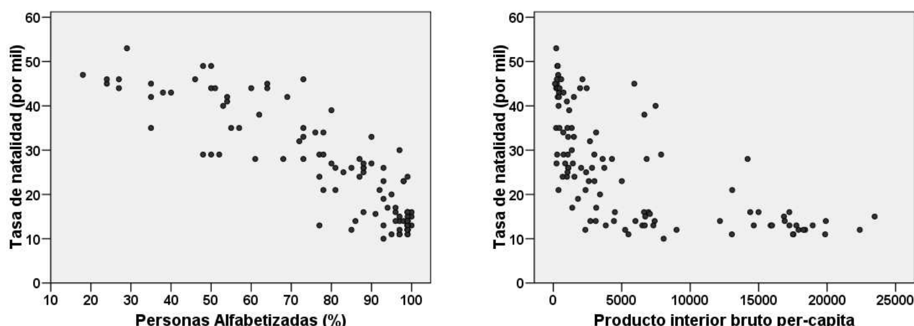
No obstante, debe tenerse en cuenta que, cuando se estudia la relación entre dos variables, no interesa únicamente poder determinar si existe o no relación lineal, sino poder cuantificar el grado o **intensidad de la relación**. Es claro que los diagramas *a* y *b* de la Figura 12.3 reflejan, ambos, una relación de tipo lineal; pero también es claro que el primero de ellos refleja una relación más fuerte o intensa que el segundo: cuanto más concentrada está la nube de puntos en torno a una línea recta, más fuerte o intensa es la relación lineal

Sin embargo, un diagrama de dispersión, por sí solo, no contiene información suficiente para poder cuantificar la intensidad de una relación de forma precisa. Entre una nube de puntos dispersa por todo el diagrama (reflejando ausencia completa de cualquier tipo de relación) y una nube de puntos concentrada en una línea recta (reflejando máximo grado de relación lineal) existen tantas y tan diversas posibles pautas de variación que hacen bastante difícil poder utilizar los diagramas de dispersión como herramientas útiles para cuantificar una relación lineal. Los diagramas de dispersión de la Figura 12.5 pueden ayudar a entender esto. Corresponden a datos reales de 109 países. El primero de ellos muestra la relación entre el porcentaje de personas alfabetizadas de cada país y la tasa de natalidad (en número de nacimientos por cada 1.000 habitantes). La nube de puntos (cada punto es un país) refleja claramente una

relación lineal negativa. Pero, ¿cómo de intensa? ¿A medio camino entre la ausencia de relación y la relación perfecta? ¿Más cerca de la ausencia de relación? ¿Más cerca de la relación perfecta? La respuesta no es nada fácil a partir de la simple inspección del diagrama.

El segundo diagrama refleja una situación aún más compleja. Representa la relación entre el producto interior bruto de cada país y la tasa de natalidad. La nube de puntos delata una relación más bien curvilínea o cuadrática, por lo que, a simple vista, no parece que la relación sea de tipo lineal. Sin embargo, lo cierto es que la relación lineal subyacente es mayor que muchas de las que suelen informarse en las ciencias sociales y de la salud como relaciones significativas y relevantes[6].

Figura 12.5. Diagramas de dispersión representando diferente grado de relación lineal



Estas consideraciones sugieren que, para poder cuantificar el grado o intensidad de una relación lineal, es necesario disponer de algún índice numérico capaz de informar de la intensidad de la relación con mayor precisión de lo que permite hacerlo la simple inspección de un diagrama de dispersión. Estos índices numéricos existen y suelen denominarse **coeficientes de correlación**. Y el más conocido y utilizado de ellos es el coeficiente de correlación de Pearson. Veamos cómo llegar a él.

Cuantificación de la intensidad de la relación: la covarianza

Del mismo modo que el centro de una variable, o su dispersión, pueden calcularse utilizando diferentes criterios (ver Capítulo 4), para cuantificar el grado de relación lineal entre dos variables también pueden seguirse distintas estrategias.

Una muy sencilla consiste en calcular en el diagrama de dispersión la proporción de puntos ubicados dentro de los cuadrantes que reflejan relación lineal. Así, en el diagrama *a* de la Figura 12.4, de los 20 puntos representados, 17 se encuentran dentro de los cuadrantes que reflejan relación lineal positiva (los cuadrantes marcados con círculos). Estos 17 puntos representan una proporción de 17/20 = 0,85. Haciendo lo mismo con los puntos del diagrama *c* se obtiene una proporción de 10/20 = 0,50. Por tanto, con esta estrategia[7] se obtienen valores

---

[6] El coeficiente de correlación de Pearson (ver siguiente apartado) entre las variables del primer diagrama vale −0,87; entre las variables del segundo diagrama vale −0,65.

[7] Aplicando esta estrategia a los diagramas de dispersión de la Figura 12.5 se obtiene una proporción de 0,89 en el primero y de 0,73 en el segundo.

(proporciones) que oscilan entre 0,50 (ausencia de relación) y 1 (máxima relación). El problema de esta estrategia es que puede ofrecer el valor máximo cuando la relación no es perfecta (ver los diagramas *XY* y *XZ* de la Figura 12.7) y el valor mínimo sin que la relación lineal sea completamente nula. Lo cual sugiere que una correcta cuantificación de la intensidad de la relación debe tener en cuenta, además del cuadrante en el que se encuentran los puntos, la ubicación concreta de los mismos dentro del cuadrante.

Esto puede conseguirse fácilmente utilizando las **puntuaciones diferenciales**. Consideremos los datos de la Tabla 12.5 referidos a las calificaciones de 8 sujetos en *lengua* (L) y *matemáticas* (M). La tabla muestra las puntuaciones directas (L y M), las diferenciales (l y m) y las medias de ambas variables (6 y 5, respectivamente). Recordemos que las puntuaciones diferenciales representan las distancias a la media:

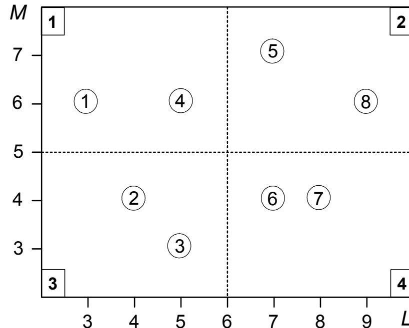
$$l = L - \overline{L}, \qquad m = M - \overline{M}, \qquad [12.8]$$

Por tanto, una puntuación diferencial *positiva* indica que la correspondiente puntuación directa es *mayor* que la media; y una puntuación diferencial *negativa* indica que la correspondiente puntuación directa es *menor* que la media. En el diagrama de dispersión de la Figura 12.6 están representados los pares de puntuaciones en *lengua* y *matemáticas* de la Tabla 12.5 (se han trazado los ejes centrales tomando como referencia la media de cada variable y se han numerado los cuadrantes resultantes para poder identificarlos fácilmente). Los puntos correspondientes a los sujetos con puntuaciones diferenciales positivas en ambas variables (sujetos 5 y 8) están ubicados en el cuadrante 2. Los puntos correspondientes a los sujetos con puntuaciones diferenciales negativas en ambas variables (sujetos 2 y 3) están ubicados en el cuadrante 3. Los puntos correspondientes a los sujetos con puntuaciones diferenciales positivas en *lengua* y negativas en *matemáticas* (sujetos 6 y 7) están ubicados en el cuadrante 4. Y los puntos correspondientes a los sujetos con puntuaciones diferenciales negativas en *lengua* y positivas en *matemáticas* (sujetos 1 y 4) están ubicados en el cuadrante 1.

Tabla 12.5. Variables L y M

Sujetos	*L*	*M*	*l*	*m*
1	3	6	−3	1
2	4	4	−2	−1
3	5	3	−1	−2
4	5	6	−1	1
5	7	7	1	2
6	7	4	1	−1
7	8	4	2	−1
8	9	6	3	1
Medias	6	5		

Figura 12.6. Diagrama de dispersión de L y M (Tabla 12.5)



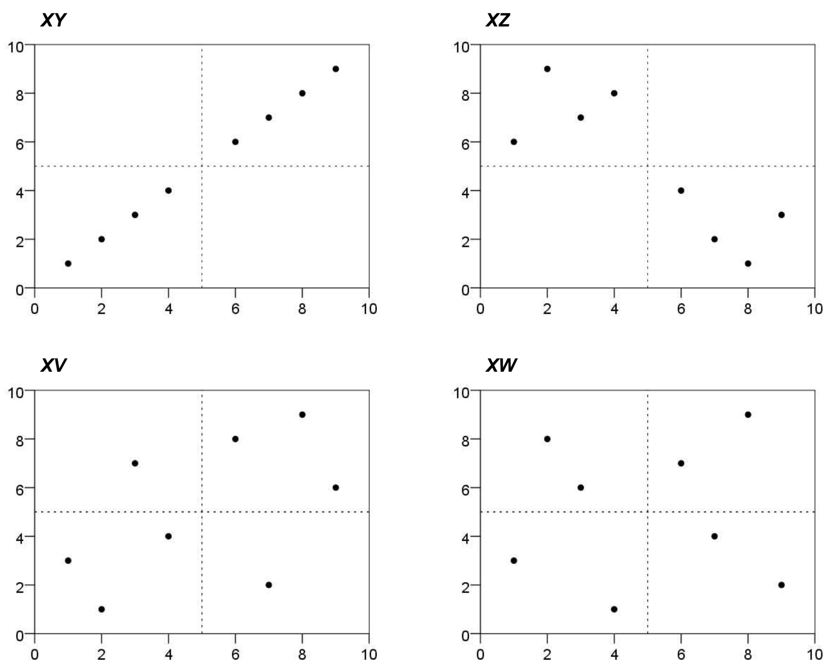
Lo interesante de esta estrategia es que las puntuaciones diferenciales, además de servir para identificar de forma precisa el cuadrante en el que se encuentra cada punto del diagrama, también sirven para indicar la distancia a la que se encuentra cada punto de cada uno de los dos ejes centrales. Por ejemplo, el punto correspondiente al primer sujeto se encuentra más alejado del eje vertical (media en *lengua*) de lo que lo está el cuarto sujeto. Y el quinto sujeto se

encuentra más alejado del eje vertical de lo que se encuentra el sexto sujeto. Por tanto, las puntuaciones diferenciales contienen toda la información necesaria para poder cuantificar el grado de relación lineal subyacente en un diagrama de dispersión: no sólo permiten identificar el cuadrante en el que se encuentra cada punto del diagrama, sino que indican con precisión la ubicación relativa de cada punto dentro su propio cuadrante.

Lo único que nos falta es acertar a operar con las puntuaciones diferenciales para que la combinación resultante pueda expresar el grado de relación. Y a Karl Pearson (1896) se le ocurrió que esto podía conseguirse simplemente multiplicando las puntuaciones diferenciales de cada par (de cada sujeto): *la relación lineal entre dos variables cuantitativas es tanto más intensa cuanto mayor es la suma de esos productos (en valor absoluto)*.

Para entender la propuesta de Pearson, consideremos los diagramas de dispersión que muestra la Figura 12.7. El diagrama *XY* revela una relación perfectamente lineal: los puntos están, todos ellos, sobre una línea recta (no podía ser de otra manera pues las variables *X* e *Y* son idénticas). El diagrama *XZ* sigue revelando una relación claramente lineal (esta vez negativa), pero menos intensa que la anterior: todos los puntos están ubicados en los cuadrantes "bajas-altas" y "altas-bajas", pero no están concentrados en una línea recta. El diagrama *XV* sigue revelando cierto grado de relación lineal, pero menos intensa aún que la anterior: aunque la mayoría de los puntos están ubicados en los cuadrantes "altas-altas" y "bajas-bajas", hay algunos puntos que se salen de esos cuadrantes. En el diagrama *XW* no hay forma de adivinar una relación lineal ni de ningún otro tipo: los puntos se encuentran dispersos por todo el diagrama, ocupando todos los cuadrantes. Por tanto, atendiendo al grado o intensidad de la relación lineal, podemos decir que ésta es mayor en *XY* que en *XZ*, mayor en *XZ* que en *XV*, y mayor en *XV* que en *XW* (*XY* > *XZ* > *XV* > *XW*).

Figura 12.7. Diagramas de dispersión correspondientes a las variables de la Tabla 12.6.



Los diagramas de dispersión de la Figura 12.7 se han elaborado a partir de la Tabla 12.6. La tabla muestra los datos de 8 sujetos en 5 variables (*X*, *Y*, *Z*, *V* y *W*). Contiene las puntuaciones directas (en mayúsculas), las diferenciales (en minúsculas) y el producto de las puntuaciones diferenciales de la variable *X* con las demás. La razón de presentar estos datos es que permiten constatar de forma clara y sencilla que la suma de los productos de las puntuaciones diferenciales (la suma de las cuatro últimas columnas de la tabla) reflejan el grado de relación existente entre las variables. Esta suma (en valor absoluto) es máxima cuando la relación es perfectamente lineal y mínima cuando no existe relación lineal: es máxima (vale 60) cuando la relación es perfecta (diagrama *XY* de la Figura 12.7); disminuye (vale 50) cuando disminuye el grado de relación lineal (diagrama *XZ*; el signo negativo de la suma indica que la relación es negativa); disminuye todavía más (vale 30) cuando sigue disminuyendo el grado de relación lineal (diagrama *XV*); y vale cero[8] (o aproximadamente cero) cuando no existe relación lineal (diagrama *XW*).

Tabla 12.6. Puntuaciones directas (en mayúsculas) y diferenciales (en minúsculas) de 5 variables (*X*, *Y*, *Z*, *V* y *W*) medidas en 8 sujetos, y productos entre las puntuaciones diferenciales de *X* y las demás

Sujetos	*X*	*Y*	*Z*	*V*	*W*	*x*	*y*	*z*	*v*	*w*	*xy*	*xz*	*xv*	*xw*
1	1	1	6	3	3	−4	−4	1	−2	−2	16	−4	8	8
2	2	2	9	1	8	−3	−3	4	−4	3	9	−12	12	−9
3	3	3	7	7	6	−2	−2	2	2	1	4	−4	−4	−2
4	4	4	8	4	1	−1	−1	3	−1	−4	1	−3	1	4
5	6	6	4	8	7	1	1	−1	3	2	1	−1	3	2
6	7	7	2	2	4	2	2	−3	−3	−1	4	−6	−6	−2
7	8	8	1	9	9	3	3	−4	4	4	9	−12	12	12
8	9	9	3	6	2	4	4	−2	1	−3	16	−8	4	−12
Sumas											60	−50	30	1

Lógicamente, esto no podía ser de otra manera pues, cuando la relación es lineal positiva (como entre *X* e *Y* en la Figura 12.7 y en Tabla 12.6), predominan los pares cuyas puntuaciones diferenciales son ambas positivas o ambas negativas; y la suma de los productos de las puntuaciones diferenciales (columna *xy*) es la suma de puntuaciones positivas. Cuando la relación es lineal negativa (como entre *XZ* en la Figura 12.7 y en Tabla 12.6), predominan los pares en los que cada puntuación diferencial es de un signo distinto; y la suma de los productos de las puntuaciones diferenciales (columna *xz*) es la suma de puntuaciones negativas. Y cuando no existe relación lineal (como entre *XW* en la Figura 12.7 y en Tabla 12.6), hay tantos pares en los que las puntuaciones diferenciales son ambas positivas o ambas negativas (pares en los que los productos son positivos) como pares en los que cada puntuación diferencial tiene un signo distinto (pares en los que los productos son negativos); y la suma de los productos de las puntuaciones diferenciales (columna *xw*) es la suma de puntuaciones positivas y negativas que tienden a anularse y sumar cero.

Ahora bien, el valor de una suma no sólo depende del valor de las puntuaciones que se suman. No es lo mismo sumar ocho puntuaciones, como en el ejemplo, que cien, o mil. Por

[8] Si no existe relación lineal, la covarianza vale cero. Pero lo contrario no es cierto. Una covarianza de cero no implica necesariamente ausencia de relación lineal; sólo implica ausencia de relación lineal entendida tal como la mide la covarianza.

tanto, la suma de los productos de las puntuaciones diferenciales no sólo depende del grado de relación lineal subyacente, sino también del número de puntuaciones sumadas. Pues bien, dividiendo esa suma entre el número de puntuaciones sumadas (no olvidar que, aquí, el número de puntuaciones se refiere al número de pares de puntuaciones; es decir, al número de productos sumados) se obtiene un estadístico muy útil y utilizado en estadística que recibe el nombre de **covarianza**[9, 10]:

$$Cov_{XY} = S_{XY} = \frac{\sum_i x_i y_i}{n-1} \qquad [12.9]$$

Aplicando la ecuación [12.9] a los datos de la Tabla 12.5 (sólo es necesario utilizar el número de casos y las sumas de las cuatro últimas columnas) se obtienen los siguientes valores:

$$S_{XY} = \frac{60}{7} = 8{,}57, \quad S_{XZ} = \frac{-50}{7} = -7{,}14, \quad S_{XV} = \frac{30}{7} = 4{,}29, \quad S_{XW} = \frac{1}{7} = 0{,}14$$

Según se ha señalado ya, el valor de la covarianza es tanto mayor (en valor absoluto) cuanto mayor es el grado de relación lineal. Y el signo de la covarianza refleja el sentido positivo o negativo de la relación. Sin embargo, como medida de la intensidad de una relación lineal, la covarianza adolece de un defecto importante: aunque su valor mínimo es cero, su valor máximo depende del grado de dispersión de las variables.

¿Qué significado tiene una covarianza S_{XY} = 8,57 o S_{XV} = 4,29? Puesto que 8,57 es el doble que 4,29 podríamos pensar que quizá la relación entre X e Y sea el doble de intensa que la relación entre X y V. Pero, al calcular la covarianza en una muestra concreta de pares de puntuaciones, ¿cómo saber si la intensidad de la relación es baja, media o alta? Por ejemplo, al multiplicar por 2 los valores de Y en la Tabla 12.5, la relación entre X e Y no se altera (es fácil comprobar que la nube de puntos del correspondiente diagrama de dispersión sigue estando sobre una línea recta); sin embargo, ese cambio en los valores de Y, que implica un cambio en su dispersión, hace que el valor de la covarianza se duplique (pasa de valer 8,57 a valer 17,14). ¿Significa este incremento en la covarianza que ha aumentado el grado de relación entre X e Y? Obviamente no, pues la relación ya era perfecta y no se ha alterado.

Que el valor mínimo de la covarianza sea cero es bueno, pues eso permite saber qué valor podemos esperar encontrar cuando no hay relación lineal. Pero que su valor máximo dependa del grado de dispersión de las variables es un problema porque eso dificulta su interpretación. No obstante, sabemos que el valor máximo de la covarianza entre dos variables es el producto de las desviaciones típicas de esas dos variables. Por tanto, el valor máximo que puede tomar la covarianza en cada situación concreta depende de las variables estudiadas y de la muestra utilizada. Esto, no sólo impide poder comparar covarianzas calculadas en distintas muestras o con distintas variables, sino que, además, complica su interpretación al tener que estar recurriendo permanentemente a máximos que dependen de cada situación.

---

[9] Se divide entre $n-1$ y no entre n por la misma razón que la varianza *insesgada* (la que siempre utilizamos por defecto) también se divide por $n-1$ (ver , en el Capítulo 7, el apartado *Estimación puntual*). El SPSS también calcula la covarianza dividiendo por $n-1$.

[10] Es evidente el paralelismo existente entre la covarianza y la varianza: si la covarianza entre X e Y se calcula sustituyendo todas las puntuaciones en Y por las puntuaciones en X, se obtiene la varianza de X; si la covarianza entre X e Y se calcula sustituyendo todas las puntuaciones en X por las puntuaciones en Y, se obtiene la varianza de Y.

El coeficiente de correlación de Pearson: R_{XY}

La solución al problema de cómo interpretar la covarianza y cómo comparar covarianzas obtenidas en muestras o variables con distinta dispersión pasa por *relativizarla* tomando como referente su valor máximo. Se obtiene así un estadístico (propuesto, también, por Karl Pearson, aunque parece que la idea fue de Galton) llamado **coeficiente de correlación de Pearson** (o *coeficiente de correlación momento-producto*):

$$R_{XY} = \frac{S_{XY}}{S_X S_Y} \qquad [12.10]$$

Esta ecuación permite interpretar R_{XY}, antes que nada, como el grado en que la covarianza alcanza su máximo. Y es equivalente a calcular la covarianza, no a partir de las puntuaciones diferenciales (las cuales reflejan la dispersión original de las puntuaciones directas), sino a partir de las puntuaciones típicas o puntuaciones Z (cuya dispersión es siempre la misma independientemente de cuál sea la dispersión de las puntuaciones directas[11]). Se obtiene de esta manera una especie de tipificación de la covarianza que ofrece idéntico resultado al de la ecuación [12.10]:

$$R_{XY} = \frac{\sum_i Z_{X_i} Z_{Y_i}}{n-1} \qquad [12.11]$$

A partir de estas dos formulaciones del coeficiente de correlación de Pearson[12] es fácil deducir sus principales **propiedades**:

- Dado que se basa en la covarianza, la magnitud de R_{XY} (en valor absoluto) *mide el grado de relación lineal* (no de otro tipo).
- Como el denominador de [12.10] siempre es positivo, *el signo de R_{XY} es el mismo que el de la covarianza*: toma un valor positivo cuando existe relación lineal positiva, un valor negativo cuando existe relación lineal negativa y un valor próximo a cero cuando no existe relación lineal.
- Puesto que se trata de la covarianza dividida entre su máximo, *su valor oscila entre* −1 *y* 1. Los valores −1 y 1 indican que la relación lineal entre las variables es perfecta.
- Al estar basado en las puntuaciones típicas (ver ecuación [12.11]), *el valor de R_{XY} no se altera si los datos se transforman linealmente* (por ejemplo, sumando y/o multiplicando una constante[13]).

---

[11] Recordemos que la varianza de las puntuaciones Z siempre vale 1 independientemente del grado de dispersión de la variable original (ver, en el Capítulo 5, el apartado *Puntuaciones Z*).

[12] Para realizar cálculos a mano puede resultar más cómodo aplicar esta otra formulación de R_{XY} a la que se llega partiendo de la ecuación [12.10] aplicando unas sencillas transformaciones:

$$R_{XY} = \frac{n\sum X_i Y_i - \sum X_i \sum Y_i}{\sqrt{n\sum X_i^2 - (\sum X_i)^2}\,\sqrt{n\sum Y_i^2 - (\sum Y_i)^2}} \qquad [12.12]$$

[13] Puesto que las trasformaciones lineales de las puntuaciones directas de una variable no alteran el valor de sus correspondientes puntuaciones típicas, la ecuación [12.11] no cambia al aplicar una transformación lineal a los datos.

Veamos cómo obtener el coeficiente de correlación de Pearson con los datos de la Tabla 12.6. Las covarianzas entre X y las demás variables las conocemos porque ya las hemos calculado al presentar la fórmula de la covarianza:

$$S_{XY} = \frac{60}{7} = 8{,}57, \quad S_{XZ} = \frac{-50}{7} = -7{,}14, \quad S_{XV} = \frac{30}{7} = 4{,}29, \quad S_{XW} = \frac{1}{7} = 0{,}14$$

Y las desviaciones típicas de las cinco variables son iguales porque en todos los casos se trata de puntuaciones de 1 a 9 en las que falta el 5. Basta, por tanto, con calcular la varianza de, por ejemplo, X. Ahora bien, como las variables X e Y son iguales, la varianza de X será idéntica a la covarianza entre X e Y. Por tanto,

$$S_X^2 = \frac{\sum_i x_i^2}{n-1} = \frac{\sum_i x_i y_i}{n-1} = \frac{60}{7} = 8{,}57 \quad \rightarrow \quad S_X = \sqrt{8{,}57} = 2{,}928$$

Aplicando ahora la ecuación [12.10] obtenemos

$$R_{XY} = \frac{8{,}57}{2{,}928 \times 2{,}928} = 1$$

$$R_{XZ} = \frac{-7{,}14}{2{,}928 \times 2{,}928} = -0{,}83$$

$$R_{XV} = \frac{4{,}29}{2{,}928 \times 2{,}928} = 0{,}50$$

$$R_{XW} = \frac{0{,}14}{2{,}928 \times 2{,}928} = 0{,}016$$

Estos valores reflejan con bastante fidelidad las pautas de variación que hemos venido discutiendo en los apartados anteriores: relación lineal positiva perfecta entre las variables X e Y ($R_{XY} = 1$; ver el diagrama XY de la Figura 12.7); relación lineal negativa y alta entre las variables X y Z ($R_{XZ} = -0{,}83$; ver el diagrama XZ de la Figura 12.7); relación lineal positiva entre las variables X y V pero claramente menos fuerte o intensa que en los dos casos anteriores ($R_{XV} = 0{,}50$; ver el diagrama XV de la Figura 12.7); y relación prácticamente nula entre las variables X y W ($R_{XW} = 0{,}016$; ver el diagrama XW de la Figura 12.7).

Contraste de hipótesis sobre el parámetro ρ_{XY}

Aplicado a unos datos concretos, el coeficiente de correlación de Pearson, R_{XY}, es un estadístico, es decir, un valor muestral. Y ya sabemos que un valor muestral, sea éste una diferencia o una correlación, puede ser distinto de cero sin que esto signifique que el correspondiente parámetro poblacional también es distinto de cero. Al comparar dos medias muestrales procedentes de la misma población o de dos poblaciones idénticas, hemos visto que una diferencia muestral podría estar reflejando simplemente las variaciones propias del azar muestral. Con un coeficiente de correlación pasa exactamente lo mismo: el hecho de que un coeficiente de correlación sea distinto de cero no constituye, en sí mismo, evidencia suficiente para afirmar que existe relación lineal en la población. Por tanto, tras cuantificar una relación, la pregunta

que hay que hacerse es si el valor muestral obtenido refleja o no un grado de parecido (grado de relación lineal) mayor del que cabría esperar por puro azar entre dos variables realmente independientes en la población.

Para responder a esta pregunta lo que suele hacerse es poner a prueba la hipótesis nula de independencia lineal (H_0: $\rho_{XY} = 0$), pues el rechazo de esta hipótesis permitirá concluir que las variables X e Y no son linealmente independientes y, por tanto, que entre ellas existe algún grado de relación lineal.

Cuando $\rho_{XY} = 0$ y las variables X e Y se distribuyen normalmente, el estadístico R_{XY} se distribuye de forma aproximadamente normal con valor esperado $E(R_{XY}) = 0$ y varianza

$$V(R_{XY}) = \hat{\sigma}^2_{R_{XY}} = \frac{1 - R^2_{XY}}{n - 2} \qquad [12.13]$$

Por tanto, la transformación resultante de restar al valor del coeficiente su valor esperado (cero) y dividir la diferencia por su error típico (raíz cuadrada de [12.13]), es decir,

$$T = \frac{R_{XY} - E(R_{XY})}{\sqrt{(1 - R^2_{XY})/(n - 2)}} = \frac{R_{XY}\sqrt{n-2}}{\sqrt{1 - R^2_{XY}}} \qquad [12.14]$$

se distribuye según el modelo de probabilidad t de Student con $n - 2$ grados de libertad. Y conociendo la distribución muestral de la trasformación [12.14], tenemos todo lo necesario para diseñar un contraste que permita poner a prueba la hipótesis nula $\rho_{XY} = 0$. El Cuadro 12.2 ofrece un resumen de este contraste.

Cuadro 12.2 Resumen del contraste sobre el parámetro ρ_{XY} (coeficiente de correlación de Pearson)[14]

1. *Hipótesis*:
   a. Contraste bilateral: H_0: $\rho_{XY} = 0$; H_1: $\rho_{XY} \neq 0$.
   b. Contraste unilateral derecho: H_0: $\rho_{XY} \leq 0$; H_1: $\rho_{XY} > 0$.
   c. Contraste unilateral izquierdo: H_0: $\rho_{XY} \geq 0$; H_1: $\rho_{XY} < 0$.
2. *Supuestos*: muestra aleatoria de n pares XY independientes entre sí procedentes de una población normal (el supuesto de normalidad va perdiendo importancia conforme va aumentando el tamaño muestral).
3. *Estadístico del contraste* (ecuación [12.14]):

$$T = \frac{R_{XY}\sqrt{n-2}}{\sqrt{1 - R^2_{XY}}}$$

[14] Esta estrategia también podría utilizarse para construir un intervalo de confianza para el parámetro ρ_{XY}. Pero no debemos olvidar que el procedimiento propuesto es válido cuando se asume que no existe relación lineal (lo cual implica $\rho_{XY} = 0$). Ahora bien, si se desea construir un intervalo de confianza para el parámetro ρ_{XY} es porque se considera que su valor es distinto de cero. Y, cuando esto ocurre, es necesario utilizar un procedimiento distinto. Ver, en el Apéndice 12, el apartado *Contraste de hipótesis sobre $\rho_{XY} = k_0$ (con $k_0 \neq 0$)*.

4. *Distribución muestral*: T se distribuye según t con $n - 2$ grados de libertad (t_{n-2}).
5. *Zona crítica*:
   a. Contraste bilateral: $T \leq t_{n-2;\ \alpha/2}$ y $T \geq t_{n-2;\ 1-\alpha/2}$.
   b. Contraste unilateral derecho: $T \geq t_{n-2;\ 1-\alpha}$.
   c. Contraste unilateral izquierdo: $T \leq t_{n-2;\ \alpha}$.
6. *Regla de decisión*: se rechaza H_0 si el estadístico del contraste cae en la zona crítica; en caso contrario, se mantiene.
   El rechazo de H_0 indica que se dispone de evidencia empírica suficiente para afirmar que, en la población, las variables X e Y están linealmente relacionadas (la forma abreviada de decir esto es que existe *relación lineal significativa*). Si no se rechaza H_0 debe concluirse que, con los datos disponibles, no es posible afirmar que X e Y estén linealmente relacionadas.
7. *Nivel crítico* (valor p):
   a. Contraste bilateral: $p = 2[P(T \geq |T_h|)]$, siendo T_h el valor muestral concreto que toma el estadístico T.
   b. Contraste unilateral derecho: $p = P(T \geq T_h)$.
   c. Contraste unilateral izquierdo: $p = P(T \leq T_h)$.

Con tamaños muestrales grandes ($n \geq 50$; ver Hays, 1994, pág. 648), en lugar de la transformación T y la distribución t puede utilizarse la distribución normal para decidir si un determinado coeficiente de correlación R_{XY} es significativamente distinto de cero. En concreto, puede rechazarse la hipótesis nula de independencia lineal cuando

$$|R_{XY}|\sqrt{n} > Z_p \qquad [12.15]$$

con $p = 1 - \alpha/2$ si el contraste es bilateral y $p = 1 - \alpha$ si el contraste es unilateral.

Ejemplo. Contraste de hipótesis sobre el parámetro ρ_{XY}

En una muestra aleatoria de 10 estudiantes de enseñanza secundaria se han medido dos variables: X = «promedio de horas de estudio semanales» e Y = «rendimiento medio» (cuantificado como la media de las calificaciones obtenidas en 8 asignaturas). Los resultados obtenidos aparecen en las primeras columnas de la Tabla 12.7. Queremos averiguar si, en la población de estudiantes de enseñanza secundaria, las puntuaciones altas en *horas de estudio* tienden a ir acompañadas de puntuaciones altas en *rendimiento medio*.

Tenemos dos variables cuantitativas medidas en una muestra aleatoria de 10 sujetos. Tenemos, por tanto, 10 pares de puntuaciones. Y queremos averiguar si, en la población de donde proceden estos 10 pares de puntuaciones, existe relación *positiva* ("... las puntuaciones altas... tienden a ir acompañadas de puntuaciones altas..."). Por tanto, podemos comenzar cuantificando el grado de relación lineal mediante el coeficiente de correlación de Pearson y, a continuación, contrastar la hipótesis nula de que las variables X e Y son linealmente independientes (en un contraste unilateral derecho).

Para facilitar los cálculos, la Tabla 12.7 recoge las puntuaciones diferenciales de ambas variables (x, y), sus cuadrados (x^2, y^2) y el producto entre ambas (xy). Estos valores se han obtenido teniendo en cuenta que la media de X vale 80/10 = 8 y la media de Y vale 60/6 = 6. Por ejemplo, los datos del primer sujeto se obtienen de la siguiente manera:

$$x = 5 - 8 = -3$$
$$y = 5 - 6 = -1$$
$$x^2 = -3^2 = 9$$
$$y^2 = -1^2 = 1$$
$$xy = -3(-1) = 3$$

Tabla 12.7. Datos de 10 sujetos en las variables X = «horas de estudio» e Y = «rendimiento medio»

Sujetos	X	Y	x	y	x^2	y^2	xy
1	5	5	−3	−1,0	9	1	3
2	5	4	−3	−2,0	9	4	6
3	6	3,5	−2	−2,5	4	6,25	5
4	6	5	−2	−1,0	4	1	2
5	6	6	−2	0	4	0	0
6	7	5	−1	−1,0	1	1	1
7	7	8	−1	2	1	4	−2,0
8	11	8,5	3	2,5	9	6,25	7,5
9	11	9	3	3	9	9	9
10	16	6	8	0	64	0	0
Totales	80	60			114	32,5	38,5

Comenzamos calculando las varianzas de X e Y y la covarianza XY para poder obtener a partir de ellas el coeficiente de correlación de Pearson:

$$S_X^2 = \frac{\sum_i x^2}{n-1} = \frac{114}{9} = 12,67 \quad \rightarrow \quad S_X = \sqrt{12,667} = 3,56$$

$$S_Y^2 = \frac{\sum_i y^2}{n-1} = \frac{32,5}{9} = 3,61 \quad \rightarrow \quad S_Y = \sqrt{3,611} = 1,90$$

$$S_{XY} = \frac{\sum_i xy}{n-1} = \frac{31,5}{9} = 3,50$$

Aplicando ahora la ecuación [12.10] se obtiene

$$R_{XY} = \frac{S_{XY}}{S_X S_Y} = \frac{3,50}{3,56\,(1,90)} = 0,52$$

Por tanto, en esta muestra, existe cierto grado de relación lineal positiva que el coeficiente de correlación de Pearson cuantifica en 0,52. Un valor intermedio que nos podría hacer sospechar que en la población existe relación lineal. Pero comprobemos si realmente esto es así:

1. *Hipótesis*: H_0: $\rho_{XY} \leq 0$.
   H_1: $\rho_{XY} > 0$ (contraste unilateral derecho: relación lineal positiva).
2. *Supuestos*: asumimos que los 10 pares de puntuaciones XY se han seleccionado aleatoriamente de una población normal.
3. *Estadístico del contraste* (ecuación [12.14]):

$$T = \frac{R_{XY}\sqrt{n-2}}{\sqrt{1-R_{XY}^2}} = \frac{0{,}52\sqrt{10-2}}{\sqrt{1-0{,}52^2}} = 1{,}72$$

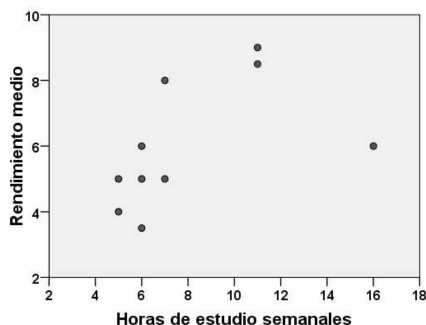
4. *Distribución muestral*: T se distribuye según t con $n-2 = 10-2 = 8$ grados de libertad.
5. Zona crítica: $T \geq t_{8;\,0{,}95} = 1{,}86$.
6. *Decisión*: como $1{,}72 < 1{,}86$, no se rechaza H_0. Por tanto, no puede afirmarse que entre las variables *horas de estudio* y *rendimiento medio* exista relación lineal positiva.
7. *Nivel crítico*: $p = P(T \geq 1{,}72) > 0{,}05$.

A pesar de que el coeficiente de correlación vale 0,52, no es posible afirmar que exista relación lineal significativa. Esto quiere decir que, al seleccionar una muestra aleatoria de 10 pares de puntuaciones XY de una población en la que las variables X e Y son linealmente independientes, el grado de relación lineal encontrado (0,52) es explicable por las fluctuaciones propias del azar muestral sin necesidad de recurrir a ninguna otra explicación.

Por otro lado, el diagrama de dispersión de la Figura 12.8 muestra un caso alejado de los demás (el par $X=16$, $Y=6$). Si se eliminara ese caso del análisis, el coeficiente de correlación subiría a 0,84 y el nivel crítico valdría 0,004, lo cual llevaría a concluir que sí existe relación lineal significativa.

Estas consideraciones nos ponen en la pista del tipo de aspectos que hay que considerar y de las precauciones que hay que tomar al interpretar un coeficiente de correlación. Al igual que cualquier otro estadístico, el coeficiente de correlación de Pearson posee sus fortalezas y debilidades y, para interpretarlo correctamente, es necesario conocerlas.

Figura 12.8. Diagrama de dispersión de *horas de estudio* por *rendimiento medio*



Cómo interpretar el coeficiente de correlación R_{XY}

El coeficiente de correlación de Pearson está, sin duda, a la cabeza de los estadísticos más utilizados para analizar la relación entre dos variables cuantitativas. Esta circunstancia podría llevar a pensar que se trata de un estadístico seguro o libre de problemas; pero nada más lejos de la realidad. Interpretar correctamente R_{XY} (o cualquier otro coeficiente de correlación) requiere prestar atención a diferentes aspectos y tomar algunas precauciones.

El coeficiente debe ser estadísticamente significativo

La primera consecuencia que se desprende del ejemplo anterior es que, antes de interpretar un coeficiente de correlación, es necesario determinar si el coeficiente es o no estadísticamente significativo. Independientemente de su valor, cuando un coeficiente no alcanza la significación estadística, no es posible afirmar que las variables están linealmente relacionadas en la población. Y, por lo general, cuando no existe evidencia de relación lineal, no tiene sentido interpretar un coeficiente.

La relevancia de un coeficiente depende del contexto

Un coeficiente estadísticamente significativo indica que existe algún grado de relación lineal. Ahora bien, esa relación, ¿cómo es de intensa? Y, ¿es positiva o negativa? Para responder a estas preguntas hay que fijarse en el valor del coeficiente y en su signo. El valor del coeficiente indica si la relación es baja, media o alta; el signo del coeficiente indica si la relación es directa (positiva) o inversa (negativa).

Por tanto, la significación estadística no lo es todo. Es más, la significación estadística de un coeficiente de correlación (al igual que la significación estadística de una diferencia entre dos medias) está estrechamente ligada al tamaño muestral. Por ejemplo, en un contraste unilateral con $\alpha = 0{,}05$ y una muestra de 20 casos, el coeficiente tiene que valer 0,37 para ser significativo; con 50 casos, 0,23; con 100 casos, 0,17; con 200 casos, 0,12; etc. Cuanto mayor es el tamaño muestral, menor necesita ser un coeficiente (en valor absoluto) para ser declarado estadísticamente significativo. Lo cual quiere decir que, con muestras grandes, se corre el riesgo de declarar estadísticamente significativos coeficientes con un valor muy próximo a cero, es decir, coeficientes que reflejan un grado de relación muy bajo.

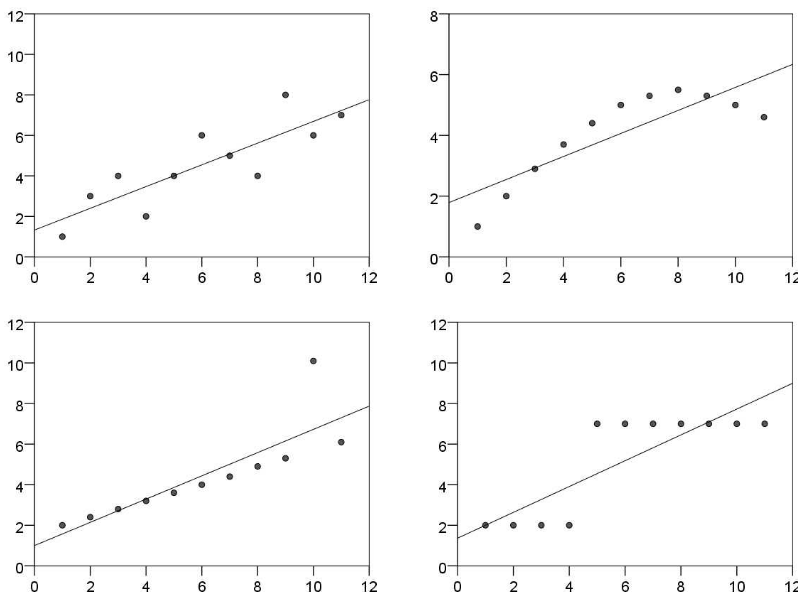
En consecuencia, tras encontrar un coeficiente de correlación es estadísticamente significativo, todavía falta por precisar si el valor del coeficiente está reflejando un grado de relación bajo, medio o alto. Y esto es algo que no puede resolver la estadística sino que, por lo general, depende del contexto. Es el propio investigador, basándose en su experiencia con un determinado tipo de variables, quien mejor puede valorar la relevancia del coeficiente de correlación encontrado. En este sentido, no es razonable adoptar reglas que indiquen cuándo un coeficiente refleja una relación baja, media o alta. Hay que referir el valor del coeficiente al contexto en el que se produce. Por ejemplo, un coeficiente de 0,70 podría corresponder a una relación poco importante o relevante cuando se trata de variables que suelen ofrecer correlaciones muy altas o cuando se obtiene en contextos donde es necesario trabajar con correlaciones muy altas (calibración de instrumentos, valoración de la fiabilidad test-retest de una escala, etc.). Y un coeficiente de 0,40 podría corresponder a una relación importante o relevante cuando se obtiene con variables que suelen correlacionar poco (variables de tipo sociodemográfico, rasgos de personalidad, etc.).

Para valorar la relevancia de un coeficiente de correlación también suele resultar bastante útil elevar el coeficiente al cuadrado. Se obtiene así un valor llamado *coeficiente de determinación* que puede interpretarse como la *proporción de varianza compartida*, es decir, como la proporción de varianza que ambas variables tienen en común (no podemos justificar esto ahora; lo haremos en el segundo volumen cuando estudiemos el análisis de regresión lineal). Por ejemplo, un coeficiente de 0,40, que en principio podría estar reflejando una relación importante en no pocos contextos, al elevarlo al cuadrado se queda en 0,16; y esto significa que las variables sólo tienen en común o comparten un 16 % de la variabilidad total.

Un mismo coeficiente puede reflejar pautas de relación muy diferentes

Consideremos los diagramas de dispersión de la Figura 12.9. El coeficiente de correlación de Pearson vale 0,84 en todos ellos. La línea recta que aparece en los diagramas atravesando la nube de puntos es un reflejo de ese valor (estudiaremos cómo trazar esa línea en el segundo volumen; de momento basta con saber que su grado de inclinación tiene que ver con el grado de relación lineal subyacente). A pesar de que la cuantificación de la relación que ofrece R_{XY} es idéntica en los cuatro casos (0,84), las nubes de puntos delatan pautas de relación muy distintas: en el primer diagrama existe una pauta claramente lineal; en el segundo, claramente cuadrática o curvilínea; en el tercero, una relación perfectamente lineal se ve alterada por la presencia de un caso anómalo; en el cuarto, las puntuaciones de la variable del eje vertical están concentradas en sólo dos valores.

Figura 12.9. Diferentes pautas de relación, todas ellas con R_{XY} = 0,84



Estos diagramas llaman la atención sobre dos cuestiones importantes. En primer lugar, un coeficiente de correlación puede resultar muy poco informativo, e incluso engañoso, si no va acompañado de una descripción apropiada de la pauta de relación subyacente. En segundo

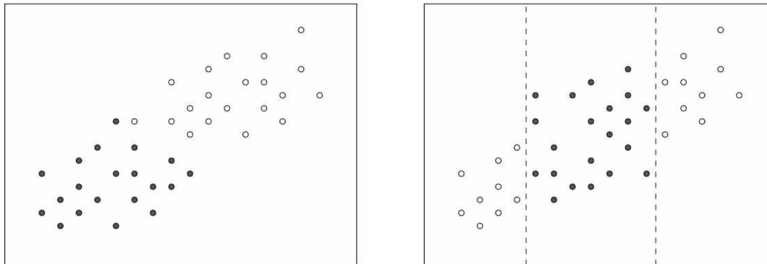
lugar, el coeficiente R_{XY} es muy sensible (poco resistente) a la presencia de casos anómalos (algo parecido ocurre en el diagrama de dispersión de la Figura 12.8); esto significa que su valor puede verse seriamente alterado por la presencia de algún caso cuyo comportamiento difiera claramente del de los demás. Estas consideraciones alertan sobre la necesidad de acompañar todo coeficiente de correlación con el correspondiente diagrama de dispersión; sólo de esa manera es posible formarse una idea acertada de lo que está ocurriendo.

Ciertas características de los datos pueden alterar de forma importante el valor de un coeficiente

La intensidad de una relación (y, consecuentemente, el valor del correspondiente coeficiente de correlación) puede verse sensiblemente alterada bajo ciertas circunstancias. Los diagramas de dispersión de la Figura 12.10 pueden ayudar a entender esto.

El primer diagrama representa la relación entre dos variables en *dos grupos* distintos (puntos negros, puntos blancos). Si se ignora la variable de agrupación (es decir, si se considera toda la nube de puntos), se obtiene un coeficiente de correlación de 0,80. Sin embargo, si se cuantifica la relación en cada grupo por separado, se obtiene un coeficiente de 0,25 para los puntos negros y de 0,41 para los puntos blancos. Esto significa que variables que están poco relacionadas pueden dar la engañosa impresión de estar muy relacionadas cuando se da la circunstancia de que un grupo (puntos blancos) tiende a dar, en ambas variables, puntuaciones sistemáticamente más altas que el otro (puntos negros). Por tanto, siempre que se tenga la sospecha de que la relación entre dos variables pueda estar modulada por terceras variables, deberán tomarse las precauciones necesarias.

Figura 12.10. Grupos distintos (izquierda) y rango de valores reducido (derecha)



Justo lo contrario (relaciones altas que pasan desapercibidas) es lo que puede ocurrir cuando no se tiene en cuenta todo el rango de posibles valores de una de las variables o de las dos. El segundo diagrama de dispersión de la Figura 12.10 muestra una situación de este tipo. Si de la variable colocada en el eje horizontal únicamente se considera el rango de valores comprendido entre las dos líneas verticales (es decir, si únicamente se tienen en cuenta los puntos negros), se puede llegar a la falsa conclusión de que las variables analizadas no están relacionadas o lo están muy poco. De hecho, si sólo se considera la nube de puntos negros se obtiene un coeficiente de correlación de 0,23. Por el contrario, si se considera todo el rango de valores del eje horizontal (es decir, toda la nube de puntos), se obtiene un coeficiente de correlación de 0,90. Por tanto, el no incluir en un análisis todo el rango de valores de las variables analizadas puede alterar sensiblemente la intensidad de la relación estudiada.

Relación y causalidad

Un coeficiente de correlación alto no implica *causalidad*. Dos variables pueden estar linealmente relacionadas (incluso muy relacionadas) sin que una sea causa de la otra.

Aunque éste no es el lugar para profundizar en el concepto de causalidad (para esto puede consultarse, por ejemplo, el excelente trabajo de Davis, 1985), sí nos parece conveniente hacer algunos comentarios al respecto. Cuando una variable es causa de otra, un cambio en la primera provoca un cambio en la segunda. Lo contrario, sin embargo, no se sostiene: cuando los cambios en una variable van asociados con cambios en la otra variable, no es posible concluir que la primera es causa de la segunda. Entre las razones de esta asimetría hay una evidente: siempre existe la posibilidad de que haya terceras variables que sean las responsables de los cambios observados en las dos variables relacionadas.

Éste es el momento de recordar las consideraciones ya hechas en el primer capítulo acerca de los diferentes niveles de indagación: las herramientas estadísticas sirven para efectuar comparaciones y estudiar relaciones, pero no dicen nada acerca de la naturaleza de las diferencias y de las relaciones encontradas. Esto depende del tipo de estrategia de recogida de datos utilizada (diseño de investigación) o de la posibilidad de contar con teorías bien fundamentadas capaces de explicar por qué ocurren las diferencias o relaciones que se observan.

Relaciones espurias

Existen innumerables ejemplos de relaciones ficticias o sin sentido que sirven para constatar que, efectivamente, una relación no implica causalidad. A este tipo de relaciones se les suele llamar *espurias* (falsas, ficticias) y no hay área de conocimiento que se libre de ellas. Veamos algunos ejemplos.

La esperanza de vida de los diferentes países está relacionada con el número de televisores por persona: la esperanza de vida tiende a ser mayor en los países con más televisores por persona. ¿Puede concluirse de esta relación que la esperanza de vida de un país depende del número de televisores de sus habitantes? ¿Regalando televisores a las personas se conseguiría alargar su vida? ¿O es más razonable pensar que debe haber alguna otra variable, como el nivel de renta, de la que depende tanto el número de televisores como la esperanza de vida y que es la verdadera responsable de la relación encontrada?

También se ha encontrado relación positiva entre el número de bomberos en un incendio y el volumen de los daños que se producen. ¿Puede concluirse de esta relación que son los bomberos quienes producen los daños? ¿Acaso reduciendo el número de bomberos se reduciría la cantidad de daños? ¿No es más razonable pensar que existe alguna otra variable, como la magnitud del incendio, de la que depende tanto el número de bomberos que acuden a sofocar el incendio como el volumen de los daños que se producen?

La altura está relacionada con el salario: las personas más altas tienden a tener mayor salario que las personas más bajas. Esta relación puede parecer sorprendente, pero es real. Quizá la explicación lógica de esta relación espuria no sea tan evidente como en los ejemplos anteriores, pero existe: el sexo actúa como variable moduladora. Tanto la altura media como el salario medio es mayor en los hombres que en las mujeres (se trata de algo parecido a lo que ocurre en el primer diagrama de la Figura 12.10, aunque con mayor solapamiento entre los puntos blancos y negros).

Las relaciones presentadas en estos tres ejemplos llaman la atención por absurdas: chocan con nuestras expectativas o ideas previas acerca de cómo son o pueden ser las cosas. No tiene

sentido para nosotros que la esperanza de vida de un país dependa del número de televisores de sus habitantes, o que los bomberos sean responsables de los daños causados por un incendio, o que el salario dependa de la altura de las personas. Son relaciones que carecen de *lógica* para nosotros y eso es lo que nos hace desconfiar de ellas.

Pero ocurre que cuando se llevan a cabo estudios reales se encuentran relaciones que, no sólo no chocan con nuestra experiencia o nuestras expectativas, sino que las refuerzan. Sin embargo, eso no quiere decir que no sean espurias. Una correlación no implica causalidad nunca; ni siquiera cuando la naturaleza de la relación nos parece del todo lógica.

Esto se entenderá mejor con un par de ejemplos menos *absurdos*. En los estudios sobre obesidad se suele encontrar una fuerte relación entre la obesidad de las madres y la de sus hijos: madres obesas tienden a tener hijos obesos. ¿Permite este dato afirmar que la obesidad de las madres es la causa, por herencia, de la obesidad de sus hijos? ¿O es posible que haya otras variables como los hábitos alimenticios o el nivel de actividad física que sean responsables de esa relación? Sólo con saber que las madres obesas tienden a tener hijos obesos no es posible afirmar que exista una relación causal entre ambas obesidades.

También suele encontrarse una relación positiva entre el nivel de estudios y el salario. ¿Se debe esto a que el salario depende del nivel de estudios? ¿O quizá a que los padres con más posibilidades económicas tienden a favorecer que sus hijos estudien y, además, a utilizar sus contactos para que sus hijos consigan mejores puestos de trabajo? Una correlación, por sí sola, no permite descartar el efecto de terceras variables y, consecuentemente, no permite afirmar la naturaleza causal de la relación subyacente.

¿Criterios de causalidad?

¿Significa lo anterior que hay que renunciar por completo a hablar de relación causa-efecto a partir de un coeficiente de correlación? En opinión de algunos autores, no del todo. Por ejemplo, Moore y Notz (2006, págs. 301-302) sostienen que, a pesar de las dificultades, existen circunstancias bajo las cuales es posible reforzar la idea de causalidad en ausencia de verdaderos experimentos. Enumeran cinco circunstancias: (1) *la relación es intensa*, por contraposición a débil; (2) *la relación es consistente*, es decir, se observa en diferentes estudios y en diferentes contextos; (3) *existe evidencia "dosis-respuesta"*, es decir, una mayor cantidad de una de las variables tiende a ir acompañada de una mayor (o menor) cantidad de la otra; (4) existe una *secuencia temporal lógica*, es decir, la variable considerada "causa" precede en el tiempo a la considerada "efecto"; y (5) *la causa alegada es plausible*, es decir, se tiene una explicación razonable de por qué las cosas son así.

Veamos si con estos criterios es posible hablar de causalidad. En el ejemplo sobre la esperanza de vida de los países y el número de televisores por persona tenemos: (1) una relación intensa, mayor que muchas de las relaciones que se informan como relevantes en las ciencias sociales y de la salud; (2) una relación consistente, pues se encuentra cada vez que se estudia el fenómeno; (3) evidencia dosis-respuesta, pues cuanto mayor es el número de televisores mayor es la esperanza de vida; (4) secuencia temporal apropiada, pues la posesión de televisores precede a la muerte; (5) no hay una explicación plausible de cómo los televisores pueden alargar la vida de las personas.

Por tanto, en una relación claramente espuria como la que se da entre la esperanza de vida y el número de televisores, se cumplen cuatro de los cinco criterios. ¿Significa esto que el hecho de que una correlación pueda interpretarse o no como causal depende de que tengamos o no una explicación plausible para ella? ¿A todos los investigadores les parecen plausibles

las mismas cosas? ¿Realmente es necesario tener una explicación plausible para todo, incluso para los nuevos hallazgos, para poder hablar de causalidad? Si la respuesta a estas preguntas es "no", entonces es claro que los cinco criterios propuestos no sirven para concluir que una relación es causal. Si la respuesta es "sí", consideremos el ejemplo sobre la relación entre el nivel de estudios y el salario. Esa relación cumple con los cinco criterios, incluido el de tener una explicación plausible (al aumentar el nivel de estudios mejora la preparación y, con ella, el acceso a puestos de mayor especialización, generalmente mejor remunerados). Sin embargo, seguimos sin poder afirmar que se trate de una relación causal porque no hemos descartado el posible efecto de terceras variables (las posibilidades económicas de los padres).

Saville (2008, págs. 114-115), basándose en las ideas de Hill (1965), añade tres nuevos criterios a los cinco anteriores (ver también Ruscio, 2006): (6) *especificidad*, es decir, si la primera variable es causa sólo de la segunda o también lo es de otras muchas; (7) *coherencia*, es decir, si la nueva relación corrobora lo que ya se sabía sobre el fenómeno o contradice resultados previos; y (8) *un cambio produce un cambio*, es decir, la manipulación de la posible "causa" produce resultados consistentes con la hipótesis causa-efecto.

¿Qué añaden estos nuevos criterios? Veamos. En primer lugar, aunque la *especificidad* puede, sin duda, ayudar a fortalecer la idea de causalidad, como criterio válido de causalidad, es poco realista. Por un lado, el hecho de que una causa no sea específica, es decir, el hecho de que sea responsable de varios efectos distintos, no significa que no sea una causa legítima (el tabaco, por ejemplo, produce problemas respiratorios, vasculares, digestivos, etc.). Por otro, el hecho de que no sepamos de qué efectos es responsable una causa no quiere decir que no lo sea de los que sí sabemos que lo es (podemos afirmar que el tabaco produce problemas respiratorios aunque no sepamos si produce o no otro tipo de problemas).

En segundo lugar, el criterio de *coherencia* también parece un criterio razonable, pues conforme se va acumulando evidencia confirmatoria sobre un fenómeno, más seguros nos vamos sintiendo de él. Sin embargo, la acumulación de conocimiento no concluyente no convierte en concluyente el conocimiento acumulado. Además, la historia de la ciencia muestra que el hallazgo de resultados no coherentes con los que ya se tienen es precisamente uno de los principales estímulos del avance del conocimiento. Y esto no parece dejar en muy buen lugar al criterio de coherencia cuando se pretende utilizar, aunque sea junto con otros, para justificar la naturaleza causal de una relación.

Por último, el criterio nº 8 implica que se está llevando a cabo un experimento. Y ya sabemos que un experimento es la principal vía para concluir causalidad. Pero en esta discusión estamos asumiendo que no hemos llevado a cabo un experimento. Estamos discutiendo bajo qué circunstancias, si las hay, una correlación permite concluir que la relación es causal. Si tenemos un experimento, sobra esta discusión. En ninguno de los ejemplos propuestos se ha utilizado un experimento. Y si el criterio nº 8 no se refiere a un experimento verdadero sino a un cuasi-experimento, entonces volvemos al punto de partida. Aunque haya manipulación de la "causa" alegada, si no hay asignación aleatoria (esto es lo que ocurre en un cuasi-experimento) no será posible descartar el efecto de terceras variables. Y si no puede descartarse el efecto de terceras variables, no es posible hablar de causalidad. Como afirma el propio promotor de los criterios de causalidad enumerados (Hill, 1965, pág. 299), para lo que sirven estos criterios es para ayudarnos a responder a la cuestión fundamental: ¿hay otra forma de explicar el hecho observado?, ¿hay otra explicación que sea tan verosímil o más que la de que la relación encontrada es causal?

Un estudio llevado a cabo por Sies (1988) ha mostrado una correlación positiva entre la tasa de nacimientos de bebés humanos y el tamaño de la población de cigüeñas alrededor de

una ciudad (el estudio se llevó a cabo en Berlín con datos relativos a la segunda mitad del siglo pasado). En un interesante trabajo, Höfer, Przyrembel y Verleger (2004), además de aportar nuevos datos en la misma dirección que los de Sies, argumentan de forma convincente que, aplicando los criterios habitualmente utilizados para extraer conclusiones causales de estudios no experimentales (práctica muy habitual en áreas donde no es posible o muy difícil la experimentación, como en epidemiología o sociología), la teoría de que a los bebés los trae la cigüeña tiene tanto o más soporte empírico que muchas otras conclusiones aceptadas sin crítica simplemente porque nos parecen *lógicas* (o no nos parecen absurdas).

La herramienta estadística utilizada en los estudios con cigüeñas es la misma que la que se utiliza en otros estudios aparentemente *serios*: una correlación. Y el tipo de diseño utilizado también es el mismo: observacional. ¿Las conclusiones de estudios esencialmente idénticos deben ser diferentes simplemente porque nos parecen más o menos absurdas? La relación entre el número de cigüeñas y la tasa de nacimientos cumple (al igual que la relación entre la esperanza de vida y el número de televisores por persona) con cuatro de los primeros cinco criterios (y ya hemos argumentado sobre la debilidad del quinto criterio). La utilidad de los criterios 6º y 7º es cuestionable. Y el 8º supone un cambio de diseño.

Una buena teoría, con predicciones precisas acerca de cómo cabe esperar que se comporten los fenómenos que se estudian, puede ayudar bastante a determinar la naturaleza de una relación (lo cual tiene que ver con lo lógica o absurda que puede parecer una relación, pero con fundamento). Ahora bien, dado que en las ciencias sociales y de la salud raramente se dan estas teorías, para poder hablar de causalidad es necesario seguir utilizando los experimentos controlados (con control sobre las condiciones del estudio y asignación aleatoria de los sujetos a esas condiciones).

Relación lineal con SPSS

Los diagramas de dispersión pueden obtenerse con la opción **Dispersión/Puntos** del **Generador de gráficos** del menú **Gráficos**. Para obtener el diagrama basta con trasladar una de las variables al eje X y otra al eje Y.

La covarianza y el coeficiente de correlación de Pearson pueden obtenerse mediante la opción **Correlaciones > Bivariadas** del menú **Analizar**. La lista de variables del cuadro de diálogo principal únicamente muestra las variables con formato numérico. Desde este cuadro de diálogo es posible obtener varios coeficientes de correlación y algunos estadísticos descriptivos básicos. La opción correspondiente al coeficiente de correlación de Pearson está marcada por defecto; para obtenerlo basta con trasladar a la lista **Variables** las variables cuyo grado de relación lineal se desea cuantificar.

Es necesario seleccionar al menos dos variables. Si se seleccionan más de dos, el SPSS calcula el coeficiente de correlación de Pearson entre cada par de variables y organiza los resultados en una matriz cuadrada con tantas filas y columnas como variables seleccionadas.

Las opciones del recuadro **Prueba de significación** permiten elegir el tipo de contraste. La opción **Bilateral** es apropiada cuando no existen expectativas sobre la *dirección* de la relación, en cuyo caso el nivel crítico indica la probabilidad de obtener coeficientes tan alejados de cero o más que el valor obtenido. La opción **Unilateral** es apropiada cuando existen expectativas sobre la *dirección* de la relación, en cuyo caso el nivel crítico indica la probabilidad de obtener coeficientes iguales o mayores que el obtenido si el coeficiente es positivo, o iguales o menores que el obtenido si el coeficiente es negativo.

La opción **Marcar las correlaciones significativas**, que se encuentra activa por defecto, hace que aparezca un asterisco al lado de los coeficientes con nivel crítico menor que 0,05 y dos asteriscos al lado de los coeficientes con nivel crítico menor que 0,01.

El subcuadro de diálogo **Opciones** permite obtener alguna información adicional (algunos estadísticos descriptivos, la covarianza, etc.) y controlar el tratamiento que se desea dar a los valores perdidos. La opción **Medias y desviaciones típicas** ofrece, para cada variable, la media, la desviación típica insesgada y el número de casos válidos. La opción **Productos cruzados y covarianzas** ofrece, para cada par de variables, la suma de los productos de las puntuaciones diferenciales y esa suma dividida por $n-1$, es decir, la covarianza.

Las opciones del recuadro **Valores perdidos** permiten elegir el tratamiento que se desea dar a los valores perdidos. La opción **Excluir casos según pareja** excluye del cálculo de cada coeficiente de correlación los casos que poseen valor perdido en alguna de las dos variables que se están correlacionando. La opción **Excluir casos según lista** excluye del cálculo de todos los coeficientes de correlación solicitados los casos que poseen valor perdido en cualquiera de las variables seleccionadas en la lista **Variables** del cuadro de diálogo principal.

Ejemplo. Relación lineal con SPSS

Este ejemplo muestra cómo obtener diagramas de dispersión, covarianzas y coeficientes de correlación de Pearson. Se basa en el archivo *Datos de empleados* (ubicado en la misma carpeta en la que está instalado el SPSS).

- ➤ Seleccionar la opción **Correlaciones > Bivariadas** del menú **Analizar** y, en el cuadro de diálogo principal, trasladar las variables *salini* (salario inicial), *salario* (salario actual) y *tiempemp* (meses desde el contrato) a la lista **Variables**.
- ➤ Pulsar el botón **Opciones** para acceder al cuadro de diálogo **Correlaciones bivariadas: Opciones** y, en el recuadro **Estadísticos**, marcar las opciones **Medias y desviaciones típicas** y **Productos cruzados y covarianzas**.

Aceptando estas elecciones, el *Visor* ofrece los resultados que muestran las Tablas 12.8 y 12.9. La Tabla 12.8 contiene únicamente información descriptiva: la media aritmética, la desviación típica insesgada y el número de casos válidos.

Tabla 12.8. Estadísticos descriptivos

	Media	Desviación típica	N
Salario actual	$34,419.57	$17,075.661	474
Salario inicial	$17,016.09	$7,870.638	474
Meses desde el contrato	81,11	10,061	474

La Tabla 12.9 ofrece información sobre el *coeficiente de correlación de Pearson* y su significación estadística. Cada celda contiene cinco valores referidos al cruce entre cada par de variables: (1) el valor del coeficiente de correlación de Pearson; (2) el nivel crítico bilateral (valor p) resultante de contrastar la hipótesis de independencia lineal, H_0: $\rho_{XY} = 0$ (*sig. bilateral*; el nivel crítico unilateral puede obtenerse dividiendo entre 2 el bilateral); (3) la suma

de los valores elevados al cuadrado (para el cruce de una variable consigo misma) o la suma de productos cruzados (para el cruce de dos variables distintas); (4) la covarianza (que se obtiene dividiendo la suma de productos cruzados entre el número de casos válidos menos uno); y (5) el número de casos válidos (*N*) que intervienen en los cálculos.

El nivel crítico (*sig*. o valor *p*) permite decidir sobre la hipótesis nula de independencia lineal (o lo que es lo mismo, sobre la hipótesis de que el coeficiente de correlación vale cero en la población). Se rechazará la hipótesis nula de independencia (y se concluirá que existe relación lineal significativa) cuando el nivel crítico sea menor que el nivel de significación establecido (generalmente, 0,05). Así, observando los niveles críticos de la Tabla 12.9, puede afirmarse que las variables *salario inicial* y *salario actual* correlacionan significativamente (*sig.* < 0,0005) y que la variable *meses desde el contrato* no correlaciona ni con la variable *salario inicial* (*sig.* = 0,668) ni con la variable *salario actual* (*sig.* = 0,067).

El SPSS no puede calcular un coeficiente de correlación cuando todos los casos de una de las variables o de ambas son casos con valor perdido, o cuando todos los casos tienen el mismo valor en una o en las dos variables correlacionadas (si todos los valores son iguales la desviación típica de esa variable vale cero). Cuando ocurre esto, el SPSS sustituye el coeficiente de correlación por una coma. Y también muestra una coma en lugar del nivel crítico (*sig.*) correspondiente al cruce de una variable consigo misma.

Tabla 12.9. Coeficientes de correlación de *Pearson* y covarianzas

		Salario actual	Salario inicial	Meses desde el contrato
Salario actual	Correlación de Pearson	1.000	.880**	.084
	Sig. (bilateral)	.	.000	.067
	Suma de cuad. y prod. cruzados	137916495436.340	55948605047.732	6833347.489
	Covarianza	291578214.453	118284577.268	14446.823
	N	474	474	474
Salario inicial	Correlación de Pearson	.880**	1.000	-.020
	Sig. (bilateral)	.000	.	.668
	Suma de cuad. y prod. cruzados	55948605047.732	29300904965.454	-739866.498
	Covarianza	118284577.268	61946944.959	-1564.200
	N	474	474	474
Meses desde el contrato	Correlación de Pearson	.084	-.020	1.000
	Sig. (bilateral)	.067	.668	.
	Suma de cuad. y prod. cruzados	6833347.489	-739866.498	47878.295
	Covarianza	14446.823	-1564.200	101.223
	N	474	474	474

** La correlación es significativa al nivel 0,01 (bilateral).

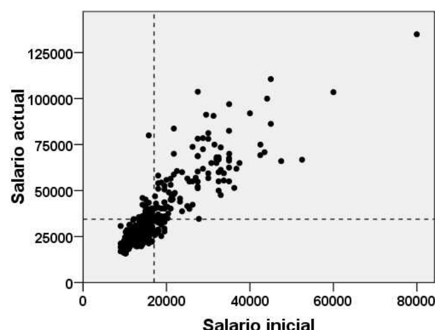
La relación entre el salario inicial y el actual (única relación estadísticamente significativa de las tres analizadas) es alta y positiva. Esto significa que las puntuaciones bajas (altas) en salario inicial tienden a ir acompañadas de puntuaciones bajas (altas) en salario actual. Pero, para formarnos una idea más acertada de esa relación, vamos a obtener el correspondiente diagrama de dispersión. Para ello:

➤ Seleccionar la opción **Generador de gráficos** del menú **Gráficos** y, en las opciones de la pestaña **Galería** del cuadro de diálogo principal, seleccionar **Dispersión/Puntos**.

➤ Trasladar la variable *salini* (salario inicial) al eje *X* y la variable *salario* (salario actual al eje *Y*.

Aceptando estas selecciones se obtiene el diagrama de dispersión que muestra la Figura 12.11. Se han trazado en el diagrama líneas punteadas a partir de las medias de cada variable. Es claro que la mayoría de los puntos están acumulados en los dos cuadrantes que corresponden a una relación lineal positiva ("bajas-bajas" y "altas-altas"). Aunque hay mayor acumulación de casos en el cuadrante "bajas-bajas" que en el cuadrante "altas-altas" (lo cual se debe a la asimetría positiva de la que adolecen ambas distribuciones), no se observan anomalías que puedan distorsionar el valor del coeficiente obtenido.

Figura 12.11. Diagrama de dispersión de *salario inicial* por *salario actual*



Apéndice 12

Contraste de hipótesis sobre $\rho_{XY} = k_0$ (con $k_0 \neq 0$)

Para contrastar la hipótesis nula de que la verdadera correlación entre dos variables es igual a un valor concreto distinto de cero (H_0: $\rho_{XY} = k_0$, con $k_0 \neq 0$) no sirve el estadístico propuesto en [12.14]. La distribución de R_{XY} se va alejando de la normalidad (se va haciendo más y más asimétrica) a medida que el valor de ρ_{XY} se va alejando de cero. No obstante, Fisher (1921) aportó una solución a este problema demostrando que la transformación[15]

$$Z_{R_{XY}} = (0,5)\log_e\left[(1+R_{XY})/(1-R_{XY})\right] \qquad [12.16]$$

se distribuye de forma aproximadamente normal con valor esperado igual al valor transformado de ρ_{XY}, es decir,

$$E(Z_{R_{XY}}) = Z_{\rho_{XY}} = (0,5)\log_e\left[(1+\rho_{XY})/(1-\rho_{XY})\right] \qquad [12.17]$$

y con varianza

$$\sigma^2_{Z_{R_{XY}}} = \frac{1}{n-3} \qquad [12.18]$$

[15] Esta transformación recibe el nombre de *zeta de Fisher* y, si se prefiere no hacer cálculos, puede consultarse la Tabla F del Apéndice final.

Esto significa que la transformación

$$Z = \frac{Z_{R_{XY}} - Z_{\rho_{XY}}}{1/\sqrt{n-3}} \qquad [12.19]$$

se distribuye de forma aproximadamente normal $N(0, 1)$. Por tanto, es posible utilizar la transformación [12.19] y la distribución normal tipificada para tomar decisiones sobre la hipótesis de que el verdadero coeficiente de correlación entre dos variables toma un valor concreto distinto de cero.

Así, por ejemplo, si en una muestra aleatoria de 20 sujetos se obtiene, para el coeficiente de correlación de Pearson, un valor de 0,65 y se desea poner a prueba la hipótesis nula de que el coeficiente de correlación poblacional vale 0,70 ($H_0: \rho_{XY} = 0{,}70$), tendremos

$$Z_{R_{XY}} = (0{,}5)\log_e\left[(1 + 0{,}65)/(1 - 0{,}65)\right] = 0{,}7753$$

$$Z_{\rho_{XY}} = (0{,}5)\log_e\left[(1 + 0{,}70)/(1 - 0{,}70)\right] = 0{,}8673$$

$$Z = \frac{Z_{R_{XY}} - Z_{\rho_{XY}}}{1/\sqrt{n-3}} = \frac{0{,}7753 - 0{,}8673}{1/\sqrt{50-3}} = -0{,}63$$

Puesto que $P(Z < -0{,}63) = 0{,}264 > \alpha$, no se rechaza H_0. La evidencia empírica disponible no permite rechazar la hipótesis de que el verdadero coeficiente de correlación vale 0,70 (es decir, los datos se muestran compatibles con la hipótesis $\rho_{XY} = 0{,}70$).

La relación establecida en [12.19] también sirve para construir un intervalo de confianza para el parámetro ρ_{XY} en los términos ya conocidos:

$$IC_{\rho_{XY}} = Z_{R_{XY}} \pm |Z_{\alpha/2}|\sqrt{n-3} \qquad [12.20]$$

Los valores obtenidos con [12.20] están en unidades de $Z_{R_{XY}}$; para obtener los límites del intervalo de confianza para ρ_{XY} hay que devolverlos a unidades de R_{XY}.

Contraste de hipótesis sobre dos coeficientes de correlación

La comparación de dos correlaciones *independientes* (dos correlaciones entre X e Y calculadas en dos muestras distintas: R_1 y R_2) es una generalización directa del procedimiento estudiado en el apartado anterior (también es posible calcular intervalos de confianza para la diferencia; ver Zou, 2007).

El estadístico $R_1 - R_2$ tiene una distribución muestral complicada de obtener, pero si se aplica la transformación de Fisher para obtener Z_{R_1} y Z_{R_2} (ecuación [12.16] o Tabla F del Apéndice final), entonces la diferencia $Z_{R_1} - Z_{R_2}$ se distribuye normalmente con valor esperado $Z_{\rho_1} - Z_{\rho_2}$ y error típico:

$$\sigma_{Z_{R_1} - Z_{R_2}} = \sqrt{1/(n_1 - 3) + 1/(n_2 - 3)} \qquad [12.21]$$

Y la transformación

$$Z = \frac{Z_{R_1} - Z_{R_2}}{\sqrt{1/(n_1 - 3) + 1/(n_2 - 3)}} \qquad [12.22]$$

se distribuye según el modelo de probabilidad normal $N(0, 1)$. Por tanto, puede utilizarse la transformación [12.22] y la distribución normal tipificada para poner a prueba la hipótesis H_0: $\rho_1 = \rho_2$ en los términos ya conocidos.

También pueden compararse coeficientes de correlación *relacionados*, es decir, calculados en la misma muestra. Tal es el caso cuando en una muestra de tamaño n se miden tres variables (X, Y y Z) y se desea averiguar, por ejemplo, si la variable X correlaciona con Y igual que con Z. Por ejemplo, en una muestra de estudiantes se miden las variables X = «rendimiento en matemáticas», Y = «aptitud numérica» y Z = «factor g» para comprobar si el *rendimiento* correlaciona más o menos con la *aptitud numérica* (R_{XY}) que con el *factor g* (R_{XZ}). Se trata de contrastar la hipótesis nula H_0: $\rho_{XY} = \rho_{XZ}$ teniendo en cuenta que los coeficientes R_{XY} y R_{XZ} se han calculado en la misma muestra.

El procedimiento tradicional utilizado para poner a prueba esta hipótesis se debe a Hotteling (1931; ver San Martín y Pardo, 1989, pág. 337). Pero Williams (1959) y, más tarde, Steiger (1980) han constatado un mejor comportamiento del estadístico

$$T = (R_{XY} - R_{XZ}) \sqrt{\frac{(n-1)(1+R_{YZ})}{2\,|R|\,\frac{n-1}{n-3} + \frac{(R_{XY}+R_{XZ})^2}{2}\,(1-R_{YZ})^3}} \qquad [12.23]$$

que se distribuye según el modelo de probabilidad t de *Student* con $n - 3$ grados de libertad. $|R|$ se refiere al determinante de la matriz de correlaciones de las tres variables y puede obtenerse fácilmente mediante

$$|R| = (1 - R_{XY}^2 - R_{XZ}^2 - R_{YZ}^2) + (2\,R_{XY}R_{XZ}R_{YZ}) \qquad [12.24]$$

Ejercicios

12.1. Antes de recibir una terapia correctora de 20 sesiones, 7 niños disléxicos han pasado por una prueba de dictado en la que se han contabilizado los errores cometidos. Tras las 20 sesiones de entrenamiento, los 7 niños han vuelto a repetir la prueba de dictado y se han vuelto a contabilizar los errores cometidos. La siguiente tabla muestra los resultados obtenidos:

Sujetos	1	2	3	4	5	6	7	$\bar{Y}_j$
Y_1: nº errores antes	19	13	20	12	15	17	9	15
Y_2: nº errores después	7	9	10	4	3	10	6	7

Utilizando $\alpha = 0{,}05$ y sabiendo que la desviación típica de las diferencias (S_{D_Y}) vale 3,606:

a. ¿Puede afirmarse que el número medio de errores ha disminuido tras el entrenamiento?

b. ¿Entre qué limites puede estimarse que se encuentra la reducción media del número de errores?

12.2. Cuando se toman dos medidas a los mismos sujetos (pre-post, o antes-después), lo que suele interesar es comparar ambas medidas para valorar si se ha producido algún cambio. Esto es lo que se ha hecho, por ejemplo, en el ejercicio anterior. Pero esto no tiene por qué ser siempre así. Ocasionalmente puede interesar constatar si el cambio observado se ha producido o

no de forma lineal, es decir, si todos los sujetos han cambiado más o menos lo mismo o de forma proporcional a sus puntuaciones originales o, por el contrario, unos sujetos han cambiado más que otros y de forma no proporcional a sus puntuaciones originales. Esto último no puede saberse comparando los promedios antes-después, sino relacionando ambas medias. Utilizando los datos del ejercicio anterior:

a. ¿Cuánto vale el coeficiente de correlación de Pearson entre los registros efectuados antes y después del entrenamiento?

b. ¿Es estadísticamente significativa la relación encontrada?

c. Explicar por qué puede haber diferencias significativas entre las mediciones antes-después y, sin embargo, no existir relación lineal significativa entre ellas.

12.3. Un investigador desea comprobar si la ingestión de alcohol reduce la capacidad de los sujetos para reconocer letras presentadas mediante taquistoscopio. Para ello, forma 10 pares aleatorios de sujetos de tal forma que los sujetos de cada par están igualados en agudeza visual. Un sujeto de cada par, seleccionado al azar, recibe una determinada dosis de alcohol. Al cabo de un tiempo preestablecido se presenta la serie de letras y se registra el número de aciertos de cada sujeto. La siguiente tabla muestra los resultados obtenidos:

Pares	1º	2º	3º	4º	5º	6º	7º	8º	9º	10º
Y_1: con alcohol	2	1	1	3	2	5	1	3	3	2
Y_2: sin alcohol	4	3	5	7	8	5	4	6	4	5

a. ¿Apoyan los datos la hipótesis de que la dosis de alcohol administrada reduce el número medio de aciertos? ($\alpha = 0{,}05$).

b. ¿Entre qué límites puede estimarse que se encuentra la diferencia en el número de aciertos con y sin alcohol? ($\alpha = 0{,}05$).

12.4. Algunos estudios sobre gemelos señalan que el miembro del par nacido en primer lugar suele mostrar un comportamiento más agresivo que el nacido en segundo lugar. Para obtener alguna evidencia más sobre esto, se ha pasado una escala de agresividad a una muestra aleatoria de 10 parejas de gemelos. La siguiente tabla muestra los resultados obtenidos:

Pares	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Y_1: 1er gemelo	23	10	15	17	22	25	20	25	11	16	13	19	21	23	10
Y_2: 2º gemelo	17	5	10	12	15	15	12	18	6	9	10	15	4	3	14

Responder a las siguientes preguntas utilizando $\alpha = 0{,}05$:

a. ¿Apoyan los datos la hipótesis de que los gemelos nacidos en primer lugar se muestran más agresivos que los nacidos en segundo lugar?

b. ¿Entre qué limites puede estimarse que se encuentra la diferencia media entre gemelos en la escala de agresividad utilizada?

c. ¿Cuánto vale el coeficiente de correlación de Pearson?

d. ¿Es estadísticamente significativa la relación encontrada?

12.5. Seguimos con los 15 pares de gemelos del ejercicio anterior. Aunque ya sabemos que el coeficiente de correlación entre las puntuaciones de los gemelos en agresividad vale 0,39, y que ese coeficiente de correlación no alcanza la significación estadística ($p > 0{,}10$), vamos a intentar formarnos una idea lo más exacta posible sobre lo que está pasando. Para ello:

a. Dibujar el correspondiente diagrama de dispersión.
b. La nube de puntos del diagrama de dispersión revela que hay tres pares de gemelos que podrían estar reduciendo sensiblemente el grado de relación lineal. ¿Cuáles son esos tres pares? Dibujar el diagrama de dispersión eliminando esos tres pares.
c. ¿Cuánto vale el coeficiente de correlación de Pearson si se eliminan esos tres pares?
d. ¿Es estadísticamente significativo el nuevo coeficiente de correlación?

12.6. Para comprobar si la intensidad luminosa (variable X) está relacionada de forma directa (positiva) con el rendimiento en una prueba de discriminación visual (variable Y), se ha utilizado el coeficiente de correlación de Pearson. Al valorar la significación del coeficiente de correlación en una muestra aleatoria de 15 sujetos se ha obtenido, para el estadístico del contraste, un valor $T = 1{,}562$. Sabiendo que $P(T \leq 1{,}562) = 0{,}93$ y utilizando un nivel de confianza de 0,99, ¿cuál de las siguientes decisiones (y motivos) es correcta?

a. Rechazar H_0 porque $P(T \leq 1{,}562) < 0{,}99$.
b. Mantener H_0 porque $P(T \leq 1{,}562) > 0{,}01$.
c. Rechazar H_0 porque $P(T \geq 1{,}562) < 0{,}01$.
d. Mantener H_0 porque $P(T \leq 1{,}562) < 0{,}99$.
e. Rechazar H_0 porque $P(T \leq 1{,}562) < 0{,}99$.

12.7. En el estudio llevado a cabo en el ejercicio anterior sobre discriminación visual se ha llegado a la conclusión de que lo razonable es no rechazar H_0. Pero:

a. ¿Qué hipótesis estadísticas se están planteando?
b. ¿Puede afirmarse que existe relación lineal entre la *intensidad luminosa* y el *rendimiento en la prueba de discriminación*? ¿Por qué?
c. ¿Cuál es el nivel de significación mínimo a partir del cual puede rechazarse H_0?

12.8. En un estudio sobre la relación entre *rigidez* y *creatividad*, un investigador plantea las siguientes hipótesis: H_0: $\rho_{XY} \geq 0$; H_1: $\rho_{XY} < 0$, y en una muestra aleatoria obtiene un estadístico $T = -2$. Sabiendo que $P(T \geq -2) = 0{,}98$ y utilizando $\alpha = 0{,}05$:

a. ¿Es razonable rechazar H_0? ¿Por qué?
b. ¿Se puede afirmar que las variables están linealmente relacionadas?

12.9. ¿Cuáles de las siguientes afirmaciones podrían servir como conclusión de los resultados obtenidos en el ejercicio anterior?

a. La creatividad no tiene nada que ver con la rigidez.
b. La creatividad depende de la rigidez.
c. La rigidez depende de la creatividad.
d. Las puntuaciones altas en rigidez tienden a ir acompañadas de puntuaciones altas en creatividad.
e. Las puntuaciones altas en rigidez tienden a ir acompañadas de puntuaciones bajas en creatividad.

12.10. Al contrastar la hipótesis nula H_0: $\rho_{XY} = 0$ frente a la alternativa H_0: $\rho_{XY} > 0$ se ha obtenido un estadístico $T = 2{,}12$. Sabiendo que $P(T > 2{,}12) = 0{,}02$, señalar la(s) alternativa(s) correcta(s):

a. Con $\alpha = 0{,}05$, se mantiene H_0.
b. Con $\alpha = 0{,}01$, puede concluirse que X está relacionada con Y.
c. Con $\alpha > 0{,}02$, se mantiene H_0.
d. Con $\alpha < 0{,}02$, se rechaza H_0.
e. Con $\alpha = 0{,}05$, puede concluirse que la correlación entre X e Y es mayor que cero.

Soluciones

12.1. *a*. Comparar dos variables cuantitativas (número de errores) medidas en los mismos sujetos. Prueba T para muestras relacionadas.

1. *Hipótesis:* $H_0 : \mu_{\text{antes}} \leq \mu_{\text{después}}$ $(\mu_D \leq 0)$.
   $H_1 : \mu_{\text{antes}} > \mu_{\text{después}}$ $(\mu_D > 0)$ (contraste unilateral derecho).
2. *Supuestos*: asumimos que la muestra de 7 diferencias se ha seleccionado aleatoriamente de una población normal.
3. *Estadístico del contraste* (ecuación [12.6]): $T = \dfrac{15-7}{3{,}606/\sqrt{7}} = \dfrac{8}{1{,}363} = 5{,}87$.
4. *Distribución muestral*: T se distribuye según t con $n-1 = 7-1 = 6$ grados de libertad.
5. *Zona crítica*: $T \geq t_{6;\,0,95} = 1{,}943$.
6. *Decisión*: como el valor del estadístico del contraste (5,87) es mayor que el punto crítico (1,943), se rechaza H_0. Por tanto, puede concluirse que el número medio de errores ha disminuido tras el entrenamiento.
7. *Nivel crítico*: $p = P(T \geq 5{,}87) < 0{,}001$.

b. Intervalo de confianza (ecuación [12.7], con $S_D/\sqrt{n} = 1{,}363$; denominador del estadístico T):

$$IC_{\mu_D} = (15-7) \pm t_{6;\,0.975}\,(1{,}363) = 8 \pm 2{,}447\,(1{,}363) = 8 \pm 3{,}34 = (4{,}66;\ 11{,}34).$$

Estimamos, con una confianza del 95%, que la verdadera disminución en el número de errores se encuentra entre 4,66 y 11,34. Teniendo en cuenta que el número de errores es una variable discreta, esto significa entre 5 y 11 errores (es decir, más de 4 y menos de 12).

12.2. *a*. Coeficiente de correlación de Pearson (ecuación [12.10]).
Comenzamos calculando las puntuaciones diferenciales para poder obtener a partir de ellas las varianzas y la covarianza:

Sujetos	1	2	3	4	5	6	7	
y_1: antes	4	−2	5	−3	0	2	−6	
y_2: después	0	2	3	−3	−4	3	−1	*Suma*
$(y_1)^2$	16	4	25	9	0	4	36	94
$(y_1)^2$	0	4	9	9	16	9	1	48
$(y_1)(y_2)$	0	−4	15	9	0	6	6	32

$S_{Y_1}^2 = \sum y_1^2/(n-1) = 94/6 = 15{,}67 \rightarrow S_{Y_1} = \sqrt{15{,}67} = 3{,}96.$

$S_{Y_2}^2 = \sum y_2^2/(n-1) = 48/6 = 8 \rightarrow S_{Y_2} = \sqrt{8} = 2{,}83.$

$S_{Y_1Y_2} = \sum y_1 y_2/(n-1) = 32/6 = 5{,}33$ (ecuación [12.9]).

$R_{XY} = S_{Y_1Y_2}/(S_{Y_1}S_{Y_2}) = 5{,}33/(3{,}96 \times 2{,}83) = 0{,}48$ (ecuación [12.10]).

b. Contraste de hipótesis sobre el coeficiente de correlación de Pearson.

1. *Hipótesis*: H_0: $\rho_{XY} = 0$.
   H_1: $\rho_{XY} \neq 0$ (contraste bilateral).
2. *Supuestos*: asumimos que los 7 pares de puntuaciones constituyen una muestra aleatoria de una población normal.
3. *Estadístico del contraste* (ecuación [12.14]): $T = \dfrac{0{,}48\sqrt{7-2}}{\sqrt{1-0{,}48^2}} = 1{,}22.$
4. *Distribución muestral*: T se distribuye según t con $n - 2 = 7 - 2 = 5$ grados de libertad (t_5).
5. *Zona crítica*: $T \leq t_{5;\,0{,}025} = -2{,}571$ y $T \geq t_{5;\,0{,}975} = 2{,}571$.
6. *Decisión*: como el valor del estadístico del contraste (1,22) se encuentra entre ambos puntos críticos, se mantiene H_0. Por tanto, no puede afirmarse que las medidas *antes* y *después* estén significativamente relacionadas.
7. *Nivel crítico*: $p = 2\,[P\,(T \geq 1{,}22)] > 0{,}20$.

c. Con variables cuantitativas, comparar y relacionar no son la misma cosa (ver el apartado *Comparar o relacionar*, al principio del capítulo). La comparación se centra en los promedios; la relación en la forma de variar las puntuaciones. El hecho de que existan diferencias quiere decir que los sujetos, en promedio, han disminuido sus errores. El hecho de que la relación entre los dos momentos no sea significativa indica que la disminución no ha sido lineal, es decir, no ha sido la misma o aproximadamente la misma en todos los sujetos, ni tampoco ha sido proporcional al número inicial de errores.

12.3. *a*. Comparar dos variables cuantitativas (número de aciertos) medidas en pares de sujetos. Prueba T para muestras relacionadas.

1. *Hipótesis:* $H_0: \mu_{con} \geq \mu_{sin}$ ($\mu_D \geq 0$).
   $H_1: \mu_{con} < \mu_{sin}$ ($\mu_D < 0$) (contraste unilateral izquierdo).

   Si el alcohol reduce la capacidad de los sujetos para reconocer palabras, el número de aciertos del grupo experimental (con alcohol) será menor que el del grupo control (sin alcohol), de ahí que se esté planteando un contraste unilateral izquierdo.
2. *Supuestos*: contamos con una muestra aleatoria de 10 diferencias extraídas de una población normal.
3. *Estadístico del contraste*:

Pares	1	2	3	4	5	6	7	8	9	10	
Y_1: Sin	2	1	1	3	2	5	1	3	3	2	
Y_2: Con	4	3	5	7	8	5	4	6	4	5	*Total*
D	−2	−2	−4	−4	−6	−0	−3	−3	−1	−3	−28
$(D - \bar{D})^2$	0,64	0,64	1,44	1,44	10,2	7,84	0	0	3,24	0	25,6

$$\bar{D} = -28/10 = -2{,}8.$$

$$S_D^2 = 25{,}6/9 = 2{,}844.$$

$$S_D = \sqrt{2{,}844} = 1{,}687.$$

$$T = \frac{-2{,}8 - 0}{1{,}687/\sqrt{10}} = \frac{-2{,}8}{0{,}533} = -5{,}25.$$

4. *Distribución muestral*: T se distribuye según t con $n - 1 = 10 - 1 = 9$ grados de libertad (t_9).
5. *Zona crítica*: $T \leq t_{9;\,0{,}05} = -1{,}833$.
6. *Decisión*: como el valor del estadístico del contraste ($-5{,}25$) es menor que el punto crítico ($-1{,}833$), se rechaza H_0. Por tanto, puede afirmarse que el promedio de aciertos del grupo experimental (con alcohol) es menor que el del grupo control (sin alcohol).
7. *Nivel crítico*: $p = P(T \leq -5{,}25) < 0{,}001$.

b. Intervalo de confianza (ecuación [12.7], con $S_D/\sqrt{n} = 0{,}533$; denominador del estadístico T):

$IC_{\mu_D} = |-2{,}8| \pm t_{9;\,0.975}\,(0{,}533) = 2{,}8 \pm 2{,}262\,(0{,}533) = 2{,}8 \pm 1{,}21 = (1{,}59;\ 4{,}01)$.

Estimamos, con una confianza del 95%, que la verdadera diferencia en el número de aciertos se encuentra entre 1,59 y 4,01. Teniendo en cuenta que el número de aciertos es una variable discreta, esto significa entre 2 y 4 aciertos (es decir, más de 1 y menos de 5).

12.4. Vamos a comenzar construyendo una tabla con los cálculos necesarios para resolver las cuestiones planteadas:

Pares	Y_1	Y_2	D	$(d)^2$	y_1	y_2	$(y_1)^2$	$(y_2)^2$	$(y_1)(y_2)$
1	23	17	6	1	5	6	25	36	30
2	10	5	5	4	−8	−6	64	36	48
3	15	10	5	4	−3	−1	9	1	3
4	17	12	5	4	−1	1	1	1	−1
5	22	15	7	0	4	4	16	16	16
6	25	15	10	9	7	4	49	16	28
7	20	12	8	1	2	1	4	1	2
8	25	18	7	0	7	7	49	49	49
9	11	6	5	4	−7	−5	49	25	35
10	16	9	7	0	−2	−2	4	4	4
11	13	10	3	16	−5	−1	25	1	5
12	19	15	4	9	1	4	1	16	4
13	21	4	17	100	3	−7	9	49	−21
14	23	3	20	169	5	−8	25	64	−40
15	10	14	−4	121	−8	3	64	9	−24
	270	165	105	442			394	324	138

$\bar{Y}_1 = 270/15 = 18 \qquad \bar{Y}_2 = 165/15 = 11 \qquad \bar{D} = 105/15 = 7.$

$S_D^2 = \sum d_Y^2/(n-1) = 442/14 = 31{,}571 \quad \rightarrow \quad S_D = \sqrt{31{,}571} = 5{,}62.$

$S_{Y_1}^2 = \sum y_1^2/(n-1) = 394/14 = 28{,}143 \quad \rightarrow \quad S_{Y_1} = \sqrt{28{,}143} = 5{,}30.$

$S_{Y_2}^2 = \sum y_2^2/(n-1) = 324/14 = 23{,}143 \quad \rightarrow \quad S_{Y_2} = \sqrt{23{,}143} = 4{,}81.$

$S_{Y_1Y_2} = \sum y_1 y_2/(n-1) = 138/14 = 9{,}86.$

a. Comparar dos variables cuantitativas (puntuaciones en agresividad) medidas en pares de sujetos. Prueba *T* para muestras relacionadas.

1. *Hipótesis:* $H_0: \mu_{antes} \leq \mu_{después}$ $(\mu_D \leq 0)$.
   $H_1: \mu_{antes} > \mu_{después}$ $(\mu_D > 0)$ (contraste unilateral derecho).
2. *Supuestos*: asumimos que la muestra de 15 diferencias se ha seleccionado aleatoriamente de una población normal.
3. *Estadístico del contraste* (ecuación [12.6]): $T = \dfrac{18-11}{5{,}62/\sqrt{15}} = \dfrac{7}{1{,}451} = 4{,}82.$
4. *Distribución muestral*: T se distribuye según t con $n-1 = 15-1 = 14$ grados de libertad.
5. *Zona crítica*: $T \geq t_{14;\,0{,}95} = 1{,}761$.
6. Decisión: como el valor del estadístico del contraste (4,82) es mayor que el punto crítico (1,761), se rechaza H_0. Por tanto, puede concluirse que la media en la escala de agresividad es mayor en los gemelos nacidos en primer lugar.
7. Nivel crítico: $p = P(T \geq 4{,}82) < 0{,}001$.

b. Intervalo de confianza (ecuación [12.7], con $S_D/\sqrt{n} = 1{,}451$; denominador del estadístico *T*):

$IC_{\mu_D} = (18-11) \pm t_{14;\,0.975}\,(1{,}451) = 7 \pm 2{,}145\,(1{,}451) = 7 \pm 3{,}11 = (3{,}89;\ 10{,}11).$

Estimamos, con una confianza del 95%, que la verdadera diferencia en las puntuaciones medias en la escala de agresividad se encuentra entre 3,89 y 10,11 puntos.

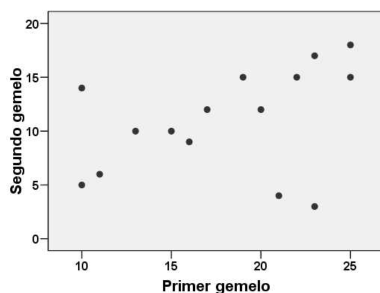
c. Coeficiente de correlación de Pearson (ecuación [12.10]):

$R_{XY} = S_{Y_1Y_2}/(S_{Y_1}S_{Y_2}) = 9{,}86/(5{,}30 \times 4{,}81) = 0{,}39.$

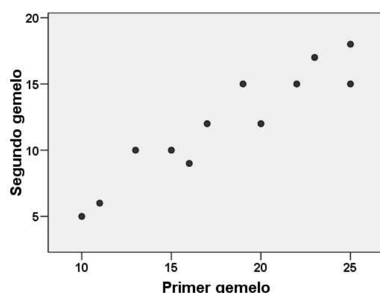
d. Contraste de hipótesis sobre el coeficiente de correlación de Pearson.

1. *Hipótesis*: H_0: $\rho_{XY} = 0$; H_1: $\rho_{XY} \neq 0$ (contraste bilateral).
2. *Supuestos*: asumimos que los 15 pares de puntuaciones constituyen una muestra aleatoria de una población normal.
3. *Estadístico del contraste* (ecuación [12.14]): $T = \dfrac{0{,}39\sqrt{15-2}}{\sqrt{1-0{,}39^2}} = 1{,}53.$
4. *Distribución muestral*: T se distribuye según t con $n-2 = 15-2 = 13$ grados de libertad (t_{13}).
5. *Zona crítica*: $T \leq t_{13;\,0{,}025} = -2{,}160$ y $T \geq t_{5;\,0{,}975} = 2{,}160$.
6. *Decisión*: como el valor del estadístico del contraste (1,53) se encuentra entre ambos puntos críticos, se mantiene H_0. Por tanto, no puede afirmarse que las puntuaciones de los gemelos en agresividad estén linealmente relacionadas.
7. *Nivel crítico*: $p = 2\,[P(T \geq 1{,}53)] > 0{,}10$.

12.5. *a.* Diagrama de dispersión:



b. El par nº 13 (con puntuaciones 21 y 4), el par nº14 (con puntuaciones 23 y 3) y el par nº15 (con puntuaciones 10 y 14). En el diagrama de dispersión se observa que a estos tres pares les corresponden puntos que se alejan de una imaginaria línea recta ascendente. En el diagrama de dispersión construido sin esos tres pares se observa claramente una nube de puntos agrupada en torno a una línea recta ascendente:



c. Coeficiente de correlación de Pearson (ecuación [12.10]) tras eliminar los pares 13, 14 y 15. Comenzamos construyendo la siguiente tabla para facilitar los cálculos:

Pares	Y_1	Y_2	y_1	y_2	$(y_1)^2$	$(y_2)^2$	$(y_1)(y_2)$
1	23	17	5	5	25	25	25
2	10	5	−8	−7	64	49	56
3	15	10	−3	−2	9	4	6
4	17	12	−1	0	1	0	0
5	22	15	4	3	16	9	12
6	25	15	7	3	49	9	21
7	20	12	2	0	4	0	0
8	25	18	7	6	49	36	42
9	11	6	−7	−6	49	36	42
10	16	9	−2	−3	4	9	6
11	13	10	−5	−2	25	4	10
12	19	15	1	3	1	9	3
	216	144			296	190	223

$$\bar{Y}_1 = 216/12 = 18, \qquad \bar{Y}_2 = 144/12 = 12.$$

$$S^2_{Y_1} = \sum y_1^2/(n-1) = 296/11 = 26{,}909 \quad \rightarrow \quad S_{Y_1} = \sqrt{26{,}909} = 5{,}19.$$

$$S^2_{Y_2} = \sum y_2^2/(n-1) = 190/11 = 17{,}273 \quad \rightarrow \quad S_{Y_2} = \sqrt{17{,}273} = 4{,}16.$$

$$S_{Y_1Y_2} = \sum y_1 y_2/(n-1) = 223/11 = 20{,}27.$$

$$R_{XY} = S_{Y_1Y_2}/(S_{Y_1}S_{Y_2}) = 20{,}27/(5{,}19 \times 4{,}16) = 0{,}94.$$

Al eliminar del análisis los pares 13, 14 y 15, el coeficiente de correlación de Pearson pasa de 0,39 a 0,94, un incremento muy importante que confirma la impresión que producen los diagramas de dispersión.

d. Contraste de hipótesis sobre el coeficiente de correlación de Pearson.

1. *Hipótesis*: H_0: $\rho_{XY} = 0$; H_1: $\rho_{XY} \neq 0$ (contraste bilateral).
2. *Supuestos*: asumimos que los 12 pares de puntuaciones constituyen una muestra aleatoria de una población normal.
3. *Estadístico del contraste* (ecuación [12.14]): $T = \dfrac{0{,}94\sqrt{12-2}}{\sqrt{1-0{,}94^2}} = 8{,}71.$
4. *Distribución muestral*: T se distribuye según t con $n - 2 = 12 - 2 = 10$ grados de libertad (t_{10}).
5. *Zona crítica*: $T \leq t_{10;\,0{,}025} = -2{,}228$ y $T \geq t_{10;\,0{,}975} = 2{,}228$.
6. *Decisión*: como el valor del estadístico del contraste (8,71) es mayor que el punto crítico derecho (2,228), se rechaza H_0. Por tanto, puede afirmarse que las puntuaciones de los gemelos en agresividad están linealmente relacionadas.
7. *Nivel crítico*: $p = 2\,[P(T \geq 8{,}71)] < 0{,}001$.

12.6. Sabemos que el contraste es unilateral derecho y que el estadístico del contraste deja por encima de sí una proporción de área igual a $1 - 0{,}93 = 0{,}07$. Por tanto, como $p = 0{,}07$ es mayor que $\alpha = 0{,}01$, la decisión razonable es rechazar H_0. Eso descarta las alternativas *a*, *c* y *e*.

La alternativa *b* es falsa porque afirma que se mantiene H_0 a partir de la comparación entre la probabilidad acumulada hasta el estadístico del contraste (que vale 0,93) y la probabilidad asociada a la zona crítica (que está a la derecha); es cierto que 0,93 es mayor que 0,01, pero esa comparación no tiene sentido. Sólo nos queda la alternativa *e*; y, aunque en ella no se está afirmando que se mantiene H_0 porque $p > \alpha$ (es decir, porque $0{,}07 > 0{,}01$), se está afirmando algo equivalente: $1 - p < 1 - \alpha$ (es decir, que $0{,}93 < 0{,}99$); y ésta es, efectivamente, la razón por la cual debe mantenerse H_0.

12.7. *a*. H_0: $\rho_{XY} \leq 0$; H_1: $\rho_{XY} > 0$.

b. No. El no rechazo de H_0 significa que no es posible afirmar que las variables estén linealmente relacionadas.

c. Nivel crítico: $p = 1 - 0{,}93 = 0{,}07$.

12.8. *a*. Sí, porque $p = P(T \leq -2) = 1 - 0{,}98 = 0{,}02 < 0{,}05$.

b. Sí. El rechazo de H_0 implica que las variables están relacionadas. Además, sabemos que la relación es negativa (contraste unilateral izquierdo).

12.9. Los resultados del ejercicio anterior indican que entre rigidez y creatividad existe relación lineal. Esto descarta la alternativa *a*. En las alternativas *b* y *c* se está afirmando que la relación detectada es de tipo

causal; y ya sabemos que para hacer este tipo de afirmaciones no basta con tomar una muestra de sujetos y medir dos variables.

Pero los resultados del ejercicio anterior no sólo indican que existe relación lineal, sino que ésta es negativa (o inversa). Esto significa que las puntuaciones altas en rigidez tienden a ir acompañadas de puntuaciones bajas en creatividad. En consecuencia, únicamente la alternativa *e* nos sirve de conclusión.

12.10. La información disponible indica que el contraste es unilateral derecho y que por encima del estadístico $T = 2{,}12$ se encuentra un área de tamaño 0,02. Esto significa que las decisiones que se tomen van a depender del nivel de significación que se adopte: un nivel de significación de 0,05 llevará a rechazar H_0; un nivel de significación de 0,01 llevará a mantener H_0.

La alternativa *a* es incorrecta, pues $0{,}02 < 0{,}05$ y eso debe llevar a rechazar H_0, no a mantenerla. Tampoco la alternativa *b* es correcta.; puesto que $0{,}02 > 0{,}01$, se debe mantener H_0 y concluir que no existe evidencia de relación lineal entre X e Y, y esto es justamente lo contrario de lo que afirma la alternativa *b*. La alternativa *c* es incorrecta por la misma razón que la *a*: cualquier nivel de significación mayor que 0,02 debe llevar a rechazar H_0, no a mantenerla. Y también la alternativa *d* es incorrecta: si α toma un valor menor que 0,02, no podrá rechazarse H_0. La única alternativa correcta es la *e*: la decisión razonable con $\alpha = 0{,}05$ es rechazar H_0 y concluir, tal como se hace en la alternativa *e*, que entre X e Y existe relación lineal positiva.

Apéndice final
Tablas estadísticas

Tabla A
Números aleatorios

	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
1	8	2	0	3	1	4	5	8	2	1	7	2	7	3	8	5	5	2	9	0	6	3	1	6	4
2	0	8	7	3	3	1	9	7	5	2	5	7	6	9	8	0	3	6	2	5	1	2	7	5	2
3	2	3	3	8	6	1	4	2	4	0	2	6	1	8	9	5	2	6	9	8	3	4	0	1	0
4	4	7	5	5	6	3	0	7	7	1	9	1	6	1	7	4	1	7	1	3	7	9	3	3	7
5	1	9	3	9	5	3	4	9	5	5	2	7	5	8	0	3	4	8	8	1	2	7	5	3	4
6	2	8	7	8	1	4	1	4	9	4	2	4	1	5	2	9	4	6	2	1	5	2	8	1	9
7	8	4	8	5	1	3	9	6	6	0	7	2	1	9	0	2	0	6	7	0	6	0	1	3	0
8	0	3	8	8	4	7	5	1	5	1	7	3	4	5	2	0	7	4	7	9	6	6	7	7	4
9	3	5	3	1	9	3	7	4	9	5	0	2	0	1	4	6	2	5	4	5	8	5	0	9	2
10	3	4	5	9	5	2	7	9	8	9	0	5	5	8	5	1	7	7	3	5	5	4	7	7	2
1	4	1	5	3	0	9	1	3	7	2	5	8	7	7	1	3	6	3	9	7	8	7	9	1	7
2	7	2	9	5	6	7	8	5	4	5	3	4	5	4	1	9	8	6	7	5	7	9	3	1	8
3	5	9	2	8	9	8	6	4	4	1	5	3	7	7	0	8	0	2	5	6	0	6	1	2	0
4	1	3	3	3	9	0	5	2	8	7	4	0	9	0	3	7	3	1	7	9	4	5	5	2	8
5	4	6	0	1	0	8	6	2	1	0	0	5	0	3	1	5	4	9	0	3	7	4	7	0	1
6	7	7	0	6	6	3	2	8	8	5	8	9	5	6	4	0	5	9	1	8	0	5	4	9	4
7	3	3	8	5	7	5	7	4	3	4	5	7	9	6	9	5	0	7	7	6	6	8	8	5	9
8	9	1	7	1	3	6	9	2	9	1	9	4	2	3	3	0	8	1	8	7	7	6	4	7	2
9	6	2	2	8	0	9	4	5	3	7	2	5	4	6	6	5	6	6	5	0	4	6	5	6	8
10	1	7	5	9	0	0	2	0	5	6	5	8	5	1	9	5	3	3	7	4	0	5	8	2	4
1	0	3	9	6	9	4	7	3	5	7	0	6	5	4	7	1	1	8	5	3	2	8	0	9	8
2	3	0	8	2	8	1	4	4	1	6	7	6	6	9	9	9	7	5	8	9	6	4	5	9	0
3	9	4	9	1	2	2	0	1	3	2	4	6	7	9	1	8	8	2	9	8	3	2	6	2	9
4	7	2	5	1	4	4	9	6	5	2	8	5	5	1	0	8	2	6	2	0	6	9	2	2	3
5	9	9	2	5	7	4	3	1	2	3	6	4	1	5	2	4	0	4	2	2	8	7	1	8	2
6	2	0	9	1	8	9	4	4	6	1	4	8	6	7	9	2	5	0	6	9	3	3	0	1	2
7	6	5	2	6	1	2	1	7	7	1	4	7	8	1	4	2	7	3	7	4	0	0	1	2	9
8	1	2	9	9	6	4	2	5	3	2	7	4	3	2	3	3	8	5	3	3	6	5	5	3	2
9	3	2	8	3	7	9	6	0	4	8	6	0	5	4	1	1	4	9	0	5	0	9	4	4	1
10	0	9	3	4	1	1	9	5	8	3	2	4	6	7	3	4	4	9	2	3	7	2	5	7	8
1	6	7	5	3	4	2	1	5	5	0	1	2	4	7	5	5	2	6	8	7	8	2	8	0	3
2	9	6	0	1	3	0	5	3	6	6	2	9	6	0	3	4	7	6	1	1	9	1	6	5	3
3	4	6	9	9	6	7	8	5	8	1	2	9	2	6	2	4	4	9	0	5	5	4	5	2	0
4	9	7	7	1	9	2	6	5	6	3	3	6	3	6	8	3	9	9	8	7	7	2	7	9	7
5	7	5	3	3	3	3	7	3	7	6	7	3	9	1	1	2	3	9	0	9	5	9	6	5	7
6	2	8	1	3	1	3	4	2	1	0	3	1	2	3	2	0	2	3	9	7	7	5	0	6	9
7	6	0	9	4	8	8	5	5	3	7	9	0	0	0	0	1	9	2	0	6	1	5	8	4	2
8	3	5	9	0	7	7	0	1	8	1	2	9	3	4	6	9	2	8	9	8	9	8	6	5	5
9	4	4	8	1	1	7	4	4	7	4	4	4	1	6	5	9	3	6	5	9	8	3	2	4	3
10	6	3	9	7	0	6	2	5	3	3	2	6	0	5	1	2	4	3	7	1	0	7	8	2	1

Tabla B

Distribuciones binomiales

Probabilidades acumuladas hasta n_1 = «número de éxitos» en cada distribución $B\,(n, \pi_1)$, con n = «número de ensayos» y π_1 = «probabilidad de éxito»

		π_1										
n	n_1	0,05	0,10	0,20	0,30	0,40	0,50	0,60	0,70	0,80	0,90	0,95
1	0	0,950	0,900	0,800	0,700	0,600	0,500	0,400	0,300	0,200	0,100	0,050
	1	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
2	0	0,902	0,810	0,640	0,490	0,360	0,250	0,160	0,090	0,040	0,010	0,002
	1	0,997	0,990	0,960	0,910	0,840	0,750	0,640	0,510	0,360	0,190	0,097
	2	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
3	0	0,857	0,729	0,512	0,343	0,216	0,125	0,064	0,027	0,008	0,001	0,000
	1	0,993	0,972	0,896	0,784	0,648	0,500	0,352	0,216	0,104	0,028	0,007
	2	1,000	0,999	0,992	0,973	0,936	0,875	0,784	0,657	0,488	0,271	0,143
	3	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
4	0	0,815	0,656	0,410	0,240	0,130	0,063	0,026	0,008	0,002	0,000	0,000
	1	0,986	0,948	0,819	0,652	0,475	0,313	0,179	0,084	0,027	0,004	0,000
	2	1,000	0,996	0,973	0,916	0,821	0,688	0,525	0,348	0,181	0,052	0,014
	3	1,000	1,000	0,998	0,992	0,974	0,938	0,870	0,760	0,590	0,344	0,185
	4	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
5	0	0,774	0,590	0,328	0,168	0,078	0,031	0,010	0,002	0,000	0,000	0,000
	1	0,977	0,919	0,737	0,528	0,337	0,188	0,087	0,031	0,007	0,000	0,000
	2	0,999	0,991	0,942	0,837	0,683	0,500	0,317	0,163	0,058	0,009	0,001
	3	1,000	1,000	0,993	0,969	0,913	0,813	0,663	0,472	0,263	0,081	0,023
	4	1,000	1,000	1,000	0,998	0,990	0,969	0,922	0,832	0,672	0,410	0,226
	5	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
6	0	0,735	0,531	0,262	0,118	0,047	0,016	0,004	0,001	0,000	0,000	0,000
	1	0,967	0,886	0,655	0,420	0,233	0,109	0,041	0,011	0,002	0,000	0,000
	2	0,998	0,984	0,901	0,744	0,544	0,344	0,179	0,070	0,017	0,001	0,000
	3	1,000	0,999	0,983	0,930	0,821	0,656	0,456	0,256	0,099	0,016	0,002
	4	1,000	1,000	0,998	0,989	0,959	0,891	0,767	0,580	0,345	0,114	0,033
	5	1,000	1,000	1,000	0,999	0,996	0,984	0,953	0,882	0,738	0,469	0,265
	6	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
7	0	0,698	0,478	0,210	0,082	0,028	0,008	0,002	0,000	0,000	0,000	0,000
	1	0,956	0,850	0,577	0,329	0,159	0,063	0,019	0,004	0,000	0,000	0,000
	2	0,996	0,974	0,852	0,647	0,420	0,227	0,096	0,029	0,005	0,000	0,000
	3	1,000	0,997	0,967	0,874	0,710	0,500	0,290	0,126	0,033	0,003	0,000
	4	1,000	1,000	0,995	0,971	0,904	0,773	0,580	0,353	0,148	0,026	0,004
	5	1,000	1,000	1,000	0,996	0,981	0,938	0,841	0,671	0,423	0,150	0,044
	6	1,000	1,000	1,000	1,000	0,998	0,992	0,972	0,918	0,790	0,522	0,302
	7	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000

Tabla B (*continuación*)

		π_1										
n	n_1	0,05	0,10	0,20	0,30	0,40	0,50	0,60	0,70	0,80	0,90	0,95
8	0	0,663	0,430	0168	0,058	0,017	0,004	0,001	0,000	0,000	0,000	0,000
	1	0,943	0,813	0,503	0,255	0,106	0,035	0,009	0,001	0,000	0,000	0,000
	2	0,994	0,962	0,797	0,552	0,315	0,145	0,050	0,011	0,001	0,000	0,000
	3	1,000	0,995	0,944	0,806	0,594	0,363	0,174	0,058	0,010	0,000	0,000
	4	1,000	1,000	0,990	0,942	0,826	0,637	0,406	0,194	0,056	0,005	0,000
	5	1,000	1,000	0,999	0,989	0,950	0,855	0,685	0,448	0,203	0,038	0,006
	6	1,000	1,000	1,000	0,999	0,991	0,965	0,894	0,745	0,497	0,187	0,057
	7	1,000	1,000	1,000	1,000	0,999	0,996	0,983	0,942	0,832	0,570	0,337
	8	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
9	0	0,630	0,387	0134	0,040	0,010	0,002	0,000	0,000	0,000	0,000	0,000
	1	0,929	0,775	0,436	0,196	0,071	0,020	0,004	0,000	0,000	0,000	0,000
	2	0,992	0,947	0,738	0,463	0,232	0,090	0,025	0,004	0,000	0,000	0,000
	3	0,999	0,992	0,914	0,730	0,483	0,254	0,099	0,025	0,003	0,000	0,000
	4	1,000	0,999	0,980	0,901	0,733	0,500	0,267	0,099	0,020	0,001	0,000
	5	1,000	1,000	0,997	0,975	0,901	0,746	0,517	0,270	0,086	0,008	0,001
	6	1,000	1,000	1,000	0,996	0,975	0,910	0,768	0,537	0,262	0,053	0,008
	7	1,000	1,000	1,000	1,000	0,996	0,980	0,929	0,804	0,564	0,225	0,071
	8	1,000	1,000	1,000	1,000	1,000	0,998	0,990	0,960	0,866	0,613	0,370
	9	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
10	0	0,599	0,349	0,107	0,028	0,006	0,001	0,000	0,000	0,000	0,000	0,000
	1	0,914	0,736	0,376	0,149	0,046	0,011	0,002	0,000	0,000	0,000	0,000
	2	0,988	0,930	0,678	0,383	0,167	0,055	0,012	0,002	0,000	0,000	0,000
	3	0,999	0,987	0,879	0,650	0,382	0,172	0,055	0,011	0,001	0,000	0,000
	4	1,000	0,998	0,967	0,850	0,633	0,377	0,166	0,047	0,006	0,000	0,000
	5	1,000	1,000	0,994	0,953	0,834	0,623	0,367	0,150	0,033	0,002	0,000
	6	1,000	1,000	0,999	0,989	0,945	0,828	0,618	0,350	0,121	0,013	0,0001
	7	1,000	1,000	1,000	0,998	0,988	0,945	0,833	0,617	0,322	0,070	0,012
	8	1,000	1,000	1,000	1,000	0,998	0,989	0,954	0,851	0,624	0,264	0,086
	9	1,000	1,000	1,000	1,000	1,000	0,999	0,994	0,972	0,893	0,651	0,401
	10	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
11	0	0,569	0,314	0,086	0,020	0,004	0,000	0,000	0,000	0,000	0,000	0,000
	1	0,898	0,697	0,322	0,113	0,030	0,006	0,001	0,000	0,000	0,000	0,000
	2	0,985	0,910	0,617	0,313	0,119	0,033	0,006	0,001	0,000	0,000	0,000
	3	0,998	0,981	0,839	0,570	0,296	0,113	0,029	0,004	0,000	0,000	0,000
	4	1,000	0,997	0,950	0,790	0,533	0,274	0,099	0,022	0,002	0,000	0,000
	5	1,000	1,000	0,988	0,922	0,753	0,500	0,247	0,078	0,012	0,000	0,000
	6	1,000	1,000	0,998	0,978	0,901	0,726	0,467	0,210	0,050	0,003	0,000
	7	1,000	1,000	1,000	0,996	0,971	0,887	0,704	0,430	0,161	0,019	0,002
	8	1,000	1,000	1,000	0,999	0,994	0,967	0,881	0,687	0,383	0,090	0,015
	9	1,000	1,000	1,000	1,000	0,999	0,994	0,970	0,887	0,678	0,303	0,102
	10	1,000	1,000	1,000	1,000	1,000	1,000	0,996	0,980	0,914	0,686	0,431
	11	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000

Tabla B (*continuación*)

		π_1										
n	n_1	0,05	0,10	0,20	0,30	0,40	0,50	0,60	0,70	0,80	0,90	0,95
12	0	0,540	0,282	0,069	0,014	0,002	0,000	0,000	0,000	0,000	0,000	0,000
	1	0,882	0,659	0,275	0,085	0,020	0,003	0,000	0,000	0,000	0,000	0,000
	2	0,980	0,889	0,558	0,253	0,083	0,019	0,003	0,000	0,000	0,000	0,000
	3	0,998	0,974	0,795	0,493	0,225	0,073	0,015	0,002	0,000	0,000	0,000
	4	1,000	0,996	0,927	0,724	0,438	0,194	0,057	0,009	0,001	0,000	0,000
	5	1,000	0,999	0,981	0,882	0,665	0,387	0,158	0,039	0,004	0,000	0,000
	6	1,000	1,000	0,996	0,961	0,842	0,613	0,335	0,118	0,019	0,001	0,000
	7	1,000	1,000	0,999	0,991	0,943	0,806	0,562	0,276	0,073	0,004	0,000
	8	1,000	1,000	1,000	0,998	0,985	0,927	0,775	0,507	0,205	0,026	0,002
	9	1,000	1,000	1,000	1,000	0,997	0,981	0,917	0,747	0,442	0,111	0,020
	10	1,000	1,000	1,000	1,000	1,000	0,997	0,980	0,915	0,725	0,341	0,118
	11	1,000	1,000	1,000	1,000	1,000	1,000	0,998	0,986	0,931	0,718	0,460
	12	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
13	0	0,513	0,254	0,055	0,010	0,001	0,000	0,000	0,000	0,000	0,000	0,000
	1	0,865	0,621	0,234	0,064	0,013	0,002	0,000	0,000	0,000	0,000	0,000
	2	0,975	0,866	0,502	0,202	0,058	0,011	0,001	0,000	0,000	0,000	0,000
	3	0,997	0,966	0,747	0,421	0,169	0,046	0,008	0,001	0,000	0,000	0,000
	4	1,000	0,994	0,901	0,654	0,353	0,133	0,032	0,004	0,000	0,000	0,000
	5	1,000	0,999	0,970	0,835	0,574	0,291	0,098	0,018	0,001	0,000	0,000
	6	1,000	1,000	0,993	0,938	0,771	0,500	0,229	0,062	0,007	0,000	0,000
	7	1,000	1,000	0,999	0,982	0,902	0,709	0,426	0,165	0,030	0,001	0,000
	8	1,000	1,000	1,000	0,996	0,968	0,867	0,647	0,346	0,099	0,006	0,000
	9	1,000	1,000	1,000	0,999	0,992	0,954	0,831	0,579	0,253	0,034	0,003
	10'	1,000	1,000	1,000	1,000	0,999	0,989	0,942	0,798	0,498	0,134	0,025
	11	1,000	1,000	1,000	1,000	1,000	0,998	0,987	0,936	0,766	0,379	0,135
	12	1,000	1,000	1,000	1,000	1,000	1,000	0,999	0,990	0,945	0,746	0,487
	13	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
14	0	0,488	0,229	0,044	0,007	0,001	0,000	0,000	0,000	0,000	0,000	0,000
	1	0,847	0,585	0,198	0,047	0,008	0,001	0,000	0,000	0,000	0,000	0,000
	2	0,970	0,842	0,448	0,161	0,040	0,006	0,001	0,000	0,000	0,000	0,000
	3	0,996	0,956	0,698	0,355	0,124	0,029	0,004	0,000	0,000	0,000	0,000
	4	1,000	0,991	0,870	0,584	0,279	0,090	0,018	0,002	0,000	0,000	0,000
	5	1,000	0,999	0,956	0,781	0,486	0,212	0,058	0,008	0,000	0,000	0,000
	6	1,000	1,000	0,988	0,907	0,692	0,395	0,150	0,031	0,002	0,000	0,000
	7	1,000	1,000	0,998	0,969	0,850	0,605	0,308	0,093	0,012	0,000	0,000
	8	1,000	1,000	1,000	0,992	0,942	0,788	0,514	0,219	0,044	0,001	0,000
	9	1,000	1,000	1,000	0,998	0,982	0,910	0,721	0,416	0,130	0,009	0,000
	10	1,000	1,000	1,000	1,000	0,996	0,971	0,876	0,645	0,302	0,044	0,004
	11	1,000	1,000	1,000	1,000	0,999	0,994	0,960	0,839	0,552	0,158	0,030
	12	1,000	1,000	1,000	1,000	1,000	0,999	0,992	0,953	0,802	0,415	0,153
	13	1,000	1,000	1,000	1,000	1,000	1,000	0,999	0,993	0,956	0,771	0,512
	14	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000

Tabla B (*continuación*)

n	n_1	π_1										
		0,05	0,10	0,20	0,30	0,40	0,50	0,60	0,70	0,80	0,90	0,95
15	0	0,463	0,206	0,035	0,005	0,000	0,000	0,000	0,000	0,000	0,000	0,000
	1	0,829	0,549	0,167	0,035	0,005	0,000	0,000	0,000	0,000	0,000	0,000
	2	0,964	0,816	0,398	0,127	0,027	0,004	0,000	0,000	0,000	0,000	0,000
	3	0,995	0,944	0,648	0,297	0,091	0,018	0,002	0,000	0,000	0,000	0,000
	4	0,999	0,987	0,836	0,515	0,217	0,059	0,009	0,001	0,000	0,000	0,000
	5	1,000	0,998	0,939	0,722	0,403	0,151	0,034	0,004	0,000	0,000	0,000
	6	1,000	1,000	0,982	0,869	0,610	0,304	0,095	0,015	0,001	0,000	0,000
	7	1,000	1,000	0,996	0,950	0,787	0,500	0,213	0,050	0,004	0,000	0,000
	8	1,000	1,000	0,999	0,985	0,905	0,696	0,390	0,131	0,018	0,000	0,000
	9	1,000	1,000	1,000	0,996	0,966	0,849	0,579	0,278	0,061	0,002	0,000
	10	1,000	1,000	1,000	0,999	0,991	0,941	0,783	0,485	0,164	0,013	0,001
	11	1,000	1,000	1,000	1,000	0,998	0,982	0,909	0,703	0,352	0,056	0,005
	12	1,000	1,000	1,000	1,000	1,000	0,996	0,973	0,873	0,602	0,184	0,036
	13	1,000	1,000	1,000	1,000	1,000	1,000	0,995	0,965	0,833	0,451	0,171
	14	1,000	1,000	1,000	1,000	1,000	1,000	1,000	0,995	0,965	0,794	0,537
	15	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
16	0	0,440	0,185	0,028	0,003	0,000	0,000	0,000	0,000	0,000	0,000	0,000
	1	0,811	0,515	0,141	0,026	0,003	0,000	0,000	0,000	0,000	0,000	0,000
	2	0,957	0,789	0,352	0,099	0,018	0,002	0,000	0,000	0,000	0,000	0,000
	3	0,993	0,932	0,598	0,246	0,065	0,011	0,001	0,000	0,000	0,000	0,000
	4	0,999	0,983	0,798	0,450	0,167	0,038	0,005	0,000	0,000	0,000	0,000
	5	1,000	0,997	0,918	0,660	0,329	0,105	0,019	0,002	0,000	0,000	0,000
	6	1,000	0,999	0,973	0,825	0,527	0,227	0,058	0,007	0,000	0,000	0,000
	7	1,000	1,000	0,993	0,926	0,716	0,402	0,142	0,026	0,001	0,000	0,000
	8	1,000	1,000	0,999	0,974	0,858	0,598	0,284	0,074	0,007	0,000	0,000
	9	1,000	1,000	1,000	0,993	0,942	0,773	0,473	0,175	0,027	0,001	0,000
	10	1,000	1,000	1,000	0,998	0,981	0,895	0,671	0,340	0,082	0,003	0,000
	11	1,000	1,000	1,000	1,000	0,995	0,962	0,833	0,550	0,202	0,017	0,001
	12	1,000	1,000	1,000	1,000	0,999	0,989	0,935	0,754	0,402	0,068	0,007
	13'	1,000	1,000	1,000	1,000	1,000	0,998	0,982	0,901	0,648	0,211	0,043
	14	1,000	1,000	1,000	1,000	1,000	1,000	0,997	0,974	0,859	0,485	0,189
	15	1,000	1,000	1,000	1,000	1,000	1,000	1,000	0,997	0,972	0,815	0,560
	16	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
17	0	0,418	0,167	0,023	0,002	0,000	0,000	0,000	0,000	0,000	0,000	0,000
	1	0,792	0,482	0,118	0,019	0,002	0,000	0,000	0,000	0,000	0,000	0,000
	2	0,950	0,762	0,310	0,077	0,012	0,001	0,000	0,000	0,000	0,000	0,000
	3	0,991	0,917	0,549	0,202	0,046	0,006	0,000	0,000	0,000	0,000	0,000
	4	0,999	0,978	0,758	0,389	0,126	0,025	0,003	0,000	0,000	0,000	0,000
	5	1,000	0,995	0,894	0,597	0,264	0,072	0,011	0,001	0,000	0,000	0,000
	6	1,000	0,999	0,962	0,775	0,448	0,166	0,035	0,003	0,000	0,000	0,000
	7	1,000	1,000	0,989	0,895	0,641	0,315	0,092	0,013	0,000	0,000	0,000
	8	1,000	1,000	0,997	0,960	0,801	0,500	0,199	0,040	0,003	0,000	0,000

Tabla B (*continuación*)

		π_1										
n	n_1	0,05	0,10	0,20	0,30	0,40	0,50	0,60	0,70	0,80	0,90	0,95
	9	1,000	1,000	1,000	0,987	0,908	0685	0359	0,105	0,011	0,000	0,000
	10	1,000	1,000	1,000	0,997	0,965	0,834	0,552	0,225	0,038	0,001	0,000
	11	1,000	1,000	1,000	0,999	0,989	0,928	0,736	0,403	0,106	0,005	0,000
	12	1,000	1,000	1,000	1,000	0,997	0,975	0,874	0,611	0,242	0,022	0,001
	13	1,000	1,000	1,000	1,000	1,000	0,994	0,954	0,798	0,451	0,083	0,009
	14	1,000	1,000	1,000	1,000	1,000	0,999	0,988	0,923	0,690	0,238	0,050
	15	1,000	1,000	1,000	1,000	1,000	1,000	0,998	0,981	0,882	0,518	0,208
	16	1,000	1,000	1,000	1,000	1,000	1,000	1,000	0,998	0,977	0,833	0,582
	17	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
18	0	0,397	0,150	0,018	0,002	0,000	0,000	0,000	0,000	0,000	0,000	0,000
	1	0,774	0,450	0,099	0,014	0,001	0,000	0,000	0,000	0,000	0,000	0,000
	2	0,942	0,734	0,271	0,060	0,008	0,001	0,000	0,000	0,000	0,000	0,000
	3	0,989	0,902	0,501	0,165	0,033	0,004	0,000	0,000	0,000	0,000	0,000
	4	0,998	0,972	0,716	0,333	0,094	0,015	0,001	0,000	0,000	0,000	0,000
	5	1,000	0,994	0,867	0,534	0,209	0,048	0,006	0,000	0,000	0,000	0,000
	6	1,000	0,999	0,949	0,722	0,374	0,119	0,020	0,001	0,000	0,000	0,000
	7	1,000	1,000	0,984	0,859	0,563	0,240	0,058	0,006	0,000	0,000	0,000
	8	1,000	1,000	0,996	0,940	0,737	0,407	0,135	0,021	0,001	0,000	0,000
	9	1,000	1,000	0,999	0,979	0,865	0,593	0,263	0,060	0,004	0,000	0,000
	10	1,000	1,000	1,000	0,994	0,942	0,760	0,437	0,141	0,016	0,000	0,000
	11	1,000	1,000	1,000	0,999	0,980	0,881	0,626	0,278	0,051	0,001	0,000
	12	1,000	1,000	1,000	1,000	0,994	0,952	0,791	0,466	0,133	0,006	0,000
	13	1,000	1,000	1,000	1,000	0,999	0,985	0,906	0,667	0,284	0,028	0,002
	14	1,000	1,000	1,000	1,000	1,000	0,996	0,967	0,835	0,499	0,098	0,011
	15'	1,000	1,000	1,000	1,000	1,000	0,999	0,992	0,940	0,729	0,266	0,058
	16	1,000	1,000	1,000	1,000	1,000	1,000	0,999	0,986	0,901	0,550	0,226
	17	1,000	1,000	1,000	1,000	1,000	1,000	1,000	0,998	0,982	0,850	0,603
	18	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
19	0	0,377	0,135	0,014	0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000
	1	0,755	0,420	0,083	0,010	0,001	0,000	0,000	0,000	0,000	0,000	0,000
	2	0,933	0,705	0,237	0,046	0,005	0,000	0,000	0,000	0,000	0,000	0,000
	3	0,987	0,885	0,455	0,133	0,023	0,002	0,000	0,000	0,000	0,000	0,000
	4	0,998	0,965	0,673	0,282	0,070	0,010	0,001	0,000	0,000	0,000	0,000
	5	1,000	0,991	0,837	0,474	0,163	0,032	0,003	0,000	0,000	0,000	0,000
	6	1,000	0,998	0,932	0,666	0,308	0,084	0,012	0,001	0,000	0,000	0,000
	7	1,000	1,000	0,977	0,818	0,488	0,180	0,035	0,003	0,000	0,000	0,000
	8	1,000	1,000	0,993	0,916	0,667	0,324	0,088	0,011	0,000	0,000	0,000
	9	1,000	1,000	0,998	0,967	0,814	0,500	0,186	0,033	0,002	0,000	0,000
	10	1,000	1,000	1,000	0,989	0,912	0,676	0,333	0,084	0,007	0,000	0,000
	11	1,000	1,000	1,000	0,997	0,965	0,820	0,512	0,182	0,023	0,000	0,000
	12	1,000	1,000	1,000	0,999	0,988	0,916	0,692	0,334	0,068	0,002	0,000
	13	1,000	1,000	1,000	1,000	0,997	0,968	0,837	0,526	0,163	0,009	0,000

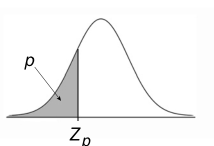
Tabla B (*continuación*)

n	n_1	π_1 0,05	0,10	0,20	0,30	0,40	0,50	0,60	0,70	0,80	0,90	0,95
	14	1,000	1,000	1,000	1,000	0,999	0,990	0,930	0,7188	0,327	0,035	0,002
	15	1,000	1,000	1,000	1,000	1,000	0,998	0,977	0,867	0,545	0,115	0,013
	16	1,000	1,000	1,000	1,000	1,000	1,000	0,995	0,954	0,763	0,295	0,067
	17	1,000	1,000	1,000	1,000	1,000	1,000	0,999	0,990	0,917	0,580	0,245
	18	1,000	1,000	1,000	1,000	1,000	1,000	1,000	0,999	0,986	0,865	0,623
	19	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000
20	0	0,358	0,122	0,012	0,001	0,000	0,000	0,000	0,000	0,000	0,000	0,000
	1	0,736	0,392	0,069	0,008	0,001	0,000	0,000	0,000	0,000	0,000	0,000
	2	0,925	0,677	0,206	0,035	0,004	0,000	0,000	0,000	0,000	0,000	0,000
	3	0,984	0,867	0,411	0,107	0,016	0,001	0,000	0,000	0,000	0,000	0,000
	4	0,997	0,957	0,630	0,238	0,051	0,006	0,000	0,000	0,000	0,000	0,000
	5	1,000	0,989	0,804	0,416	0,126	0,021	0,002	0,000	0,000	0,000	0,000
	6	1,000	0,998	0,913	0,608	0,250	0,058	0,006	0,000	0,000	0,000	0,000
	7	1,000	1,000	0,968	0,772	0,416	0,132	0,021	0,001	0,000	0,000	0,000
	8	1,000	1,000	0,990	0,887	0,596	0,252	0,057	0,005	0,000	0,000	0,000
	9	1,000	1,000	0,997	0,952	0,755	0,412	0,128	0,017	0,001	0,000	0,000
	10	1,000	1,000	0,999	0,983	0,872	0,588	0,245	0,048	0,003	0,000	0,000
	11	1,000	1,000	1,000	0,995	0,943	0,748	0,404	0,113	0,010	0,000	0,000
	12	1,000	1,000	1,000	0,999	0,979	0,868	0,584	0,228	0,032	0,000	0,000
	13	1,000	1,000	1,000	1,000	0,994	0,942	0,750	0,392	0,087	0,002	0,000
	14	1,000	1,000	1,000	1,000	0,998	0,979	0,874	0,584	0,196	0,011	0,000
	15	1,000	1,000	1,000	1,000	1,000	0,994	0,949	0,762	0,370	0,043	0,003
	16	1,000	1,000	1,000	1,000	1,000	0,999	0,984	0,893	0,589	0,133	0,016
	17'	1,000	1,000	1,000	1,000	1,000	1,000	0,996	0,965	0,794	0,323	0,075
	18	1,000	1,000	1,000	1,000	1,000	1,000	0,999	0,992	0,931	0,608	0,264
	19	1,000	1,000	1,000	1,000	1,000	1,000	1,000	0,999	0,988	0,878	0,642
	20	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000

Tabla C

Distribución normal tipificada: $N(0, 1)$

Probabilidades acumuladas (p) hasta cada valor Z



Z_p	Segundo decimal de Z_p									
	0	1	2	3	4	5	6	7	8	9
−3,2	0,0007	0,0007	0,0006	0,0006	0,0006	0,0006	0,0006	0,0005	0,0005	0,0005
−3,1	0,0010	0,0009	0,0009	0,0009	0,0008	0,0008	0,0008	0,0008	0,0007	0,0007
−3,0	0,0013	0,0013	0,0013	0,0012	0,0012	0,0012	0,0011	0,0011	0,0010	0,0010
−2,9	0,0019	0,0018	0,0017	0,0017	0,0016	0,0016	0,0015	0,0015	0,0014	0,0014
−2,8	0,0026	0,0025	0,0024	0,0023	0,0023	0,0022	0,0021	0,0020	0,0020	0,0019
−2,7	0,0035	0,0034	0,0033	0,0032	0,0031	0,0030	0,0029	0,0028	0,0027	0,0026
−2,6	0,0047	0,0045	0,0044	0,0043	0,0041	0,0040	0,0039	0,0038	0,0037	0,0036
−2,5	0,0062	0,0060	0,0059	0,0057	0,0055	0,0054	0,0052	0,0051	0,0049	0,0048
−2,4	0,0082	0,0080	0,0078	0,0075	0,0073	0,0071	0,0069	0,0068	0,0066	0,0064
−2,3	0,0107	0,0104	0,0102	0,0099	0,0096	0,0094	0,0091	0,0089	0,0087	0,0084
−2,2	0,0139	0,0136	0,0132	0,0129	0,0125	0,0122	0,0119	0,0116	0,0113	0,0110
−2,1	0,0179	0,0174	0,0170	0,0166	0,0162	0,0158	0,0154	0,0150	0,0146	0,0143
−2,0	0,0228	0,0222	0,0217	0,0212	0,0207	0,0202	0,0197	0,0192	0,0188	0,0183
−1,9	0,0287	0,0281	0,0274	0,0268	0,0262	0,0256	0,0250	0,0244	0,0239	0,0233
−1,8	0,0359	0,0351	0,0344	0,0336	0,0329	0,0322	0,0314	0,0307	0,0301	0,0294
−1,7	0,0446	0,0436	0,0427	0,0418	0,0409	0,0401	0,0392	0,0384	0,0375	0,0367
−1,6	0,0548	0,0537	0,0526	0,0516	0,0505	0,0495	0,0485	0,0475	0,0465	0,0455
−1,5	0,0668	0,0655	0,0643	0,0630	0,0618	0,0606	0,0594	0,0582	0,0571	0,0559
−1,4	0,0808	0,0793	0,0778	0,0764	0,0749	0,0735	0,0721	0,0708	0,0694	0,0681
−1,3	0,0968	0,0951	0,0934	0,0918	0,0901	0,0885	0,0869	0,0853	0,0838	0,0823
−1,2	0,1151	0,1131	0,1112	0,1093	0,1075	0,1056	0,1038	0,1020	0,1003	0,0985
−1,1	0,1357	0,1335	0,1314	0,1292	0,1271	0,1251	0,1230	0,1210	0,1190	0,1170
−1,0	0,1587	0,1562	0,1539	0,1515	0,1492	0,1469	0,1446	0,1423	0,1401	0,1379
−0,9	0,1841	0,1814	0,1788	0,1762	0,1736	0,1711	0,1685	0,1660	0,1635	0,1611
−0,8	0,2119	0,2090	0,2061	0,2033	0,2005	0,1977	0,1949	0,1922	0,1894	0,1867
−0,7	0,2420	0,2389	0,2358	0,2327	0,2296	0,2266	0,2236	0,2206	0,2177	0,2148
−0,6	0,2743	0,2709	0,2676	0,2643	0,2611	0,2578	0,2546	0,2514	0,2483	0,2451
−0,5	0,3085	0,3050	0,3015	0,2981	0,2946	0,2912	0,2877	0,2843	0,2810	0,2776
−0,4	0,3446	0,3409	0,3372	0,3336	0,3300	0,3264	0,3228	0,3192	0,3156	0,3121
−0,3	0,3821	0,3783	0,3745	0,3707	0,3669	0,3632	0,3594	0,3557	0,3520	0,3483
−0,2	0,4207	0,4168	0,4129	0,4090	0,4052	0,4013	0,3974	0,3936	0,3897	0,3859
−0,1	0,4602	0,4562	0,4522	0,4483	0,4443	0,4404	0,4364	0,4325	0,4286	0,4247
−0,0	0,5000	0,4960	0,4920	0,4880	0,4840	0,4801	0,4761	0,4721	0,4681	0,4641

Tabla C (*continuación*)

Z_p	Segundo decimal de Z_p									
	0	1	2	3	4	5	6	7	8	9
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,0	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990
3,1	0,9990	0,9991	0,9991	0,9991	0,9992	0,9992	0,9992	0,9992	0,9993	0,9993
3,2	0,9993	0,9993	0,9994	0,9994	0,9994	0,9994	0,9994	0,9995	0,9995	0,9995

Valores Z_p seleccionados: $Z_{0,90} = 1,282$ $Z_{0,95} = 1,645$ $Z_{0,975} = 1,960$

$Z_{0,99} = 2,326$ $Z_{0,995} = 2,576$ $Z_{0,999} = 3,090$

$Z_{0,9994} = 3,25$ $Z_{0,9998} = 3,50$ $Z_{0,9999} = 3,75$

Tabla D

Distribuciones χ^2 (*ji*-cuadrado)

Valores $\chi^2_{gl;\,p}$ que acumulan una probabilidad p con diferentes grados de libertad (gl)

	p = probabilidad acumulada hasta el valor $\chi^2_{gl;\,p}$											
gl	0,001	0,005	0,01	0,025	0,050	0,100	0,900	0,950	0,975	0,990	0,995	0,999
1	0,00	0,00	0,00	0,00	0,00	0,02	2,71	3,84	5,02	6,63	7,88	10,83
2	0,00	0,01	0,02	0,05	0,10	0,21	4,61	5,99	7,38	9,21	10,60	13,82
3	0,02	0,07	0,12	0,22	0,35	0,58	6,25	7,81	9,35	11,34	12,84	16,27
4	0,09	0,21	0,28	0,48	0,71	1,06	7,78	9,49	11,14	13,28	14,86	18,47
5	0,21	0,41	0,55	0,83	1,14	1,61	9,24	11,07	12,83	15,09	16,75	20,52
6	0,38	0,68	0,87	1,24	1,64	2,20	10,64	12,59	14,45	16,81	18,55	22,46
7	0,60	0,99	1,24	1,69	2,27	2,83	12,02	14,07	16,01	18,48	20,28	24,32
8	0,86	1,34	1,65	2,18	2,73	3,49	13,36	15,51	17,53	20,09	21,96	26,13
9	1,15	1,73	2,09	2,70	3,33	4,17	14,68	16,92	19,02	21,67	23,59	27,88
10	1,48	2,16	2,56	3,25	3,94	4,87	15,99	18,31	20,48	23,21	25,19	29,59
11	1,83	2,60	3,05	3,82	4,57	5,58	17,28	19,68	21,92	24,72	26,76	31,26
12	2,21	3,07	3,57	4,40	5,23	6,30	18,55	21,03	23,34	26,22	28,30	32,91
13	2,62	3,57	4,11	5,01	5,89	7,04	19,81	22,36	24,74	27,69	29,82	34,53
14	3,04	4,07	4,66	5,63	6,57	7,79	21,06	23,68	26,12	29,14	31,32	36,12
15	3,48	4,60	5,23	6,26	7,26	8,55	22,31	25,00	27,49	30,58	32,80	37,70
16	3,94	5,14	5,81	6,91	7,96	9,31	23,54	26,30	28,85	32,00	34,27	39,25
17	4,42	5,70	6,41	7,56	8,67	10,09	24,77	27,59	30,19	33,41	35,72	40,79
18	4,90	6,26	7,01	8,23	9,39	10,86	25,99	28,87	31,53	34,81	37,16	42,31
19	5,41	6,84	7,63	8,91	10,12	11,65	27,20	30,14	32,85	36,19	38,58	43,82
20	5,92	7,43	8,26	8,59	10,85	12,44	28,41	31,41	34,17	37,57	40,00	45,31
21	6,45	8,03	8,90	10,28	11,59	13,24	29,62	32,67	35,48	38,93	41,40	46,80
22	6,98	8,64	9,54	10,98	12,34	14,04	30,81	33,92	36,78	40,29	42,80	48,27
23	7,53	9,26	10,20	11,69	13,09	14,85	32,01	35,17	38,08	41,64	44,18	49,73
24	8,08	9,89	10,86	12,40	13,85	15,66	33,20	36,42	39,36	42,98	45,56	51,18
25	8,65	10,52	11,52	13,12	14,61	16,47	34,38	37,65	40,65	44,31	46,93	52,62
26	9,22	11,16	12,20	13,84	15,38	17,29	35,56	38,89	41,92	45,64	48,29	54,05
27	9,80	11,81	12,88	14,57	16,15	18,11	36,74	40,11	43,19	46,96	49,64	55,48
28	10,39	12,46	13,56	15,31	16,39	18,94	37,92	41,34	44,46	48,28	50,99	56,89
29	10,99	13,21	14,26	16,05	17,71	19,77	39,09	42,56	45,72	49,59	52,34	58,30
30	11,59	13,79	14,95	16,79	18,49	20,60	40,26	43,77	46,98	50,89	53,67	59,70
40	17,92	20,71	22,16	24,43	26,51	29,05	51,81	55,76	59,34	63,69	66,77	73,40
50	24,67	27,99	29,71	32,36	34,76	37,69	63,17	67,50	71,42	76,15	79,49	86,66
60	31,74	35,53	37,48	40,48	43,19	46,46	74,40	79,08	83,30	88,38	91,95	99,61
70	39,04	43,28	45,44	48,76	51,74	55,33	85,53	90,53	95,02	100,43	104,21	112,32
80	46,52	51,17	53,54	57,15	60,39	64,28	96,58	101,88	106,63	112,33	116,32	124,84
90	54,16	59,20	61,75	65,65	69,13	73,29	107,57	113,15	118,14	124,12	128,30	137,21
100	61,92	67,33	70,06	74,22	77,93	82,36	118,50	124,34	129,56	135,81	140,17	149,45

Con $gl > 30$, puede utilizarse la aproximación: $\chi^2_{gl;\,p} \approx \frac{1}{2}\,[Z_p + \sqrt{2gl-1}\,]^2$

Tabla E

Distribuciones *t* de Student

Valores $t_{gl;\,p}$ que acumulan una probabilidad p con diferentes grados de libertad gl $(t_{gl;\,p} = -t_{gl;\,1-p})$

	p = probabilidad acumulada hasta el valor $t_{gl;\,p}$											
gl	0,001	0,005	0,010	0,025	0,050	0,100	0,900	0,950	0,975	0,990	0,995	0,999
2	−22,32	−9,925	−6,965	−2,920	−2,920	−1,886	1,886	2,920	4,303	6,965	9,925	22,32
3	−10,21	−5,841	−4,541	−2,353	−2,353	−1,638	1,638	2,353	3,182	4,541	5,841	10,21
4	−7,173	−4,604	−3,747	−2,132	−2,132	−1,533	1,533	2,132	2,776	3,747	4,604	7,173
5	−5,893	−4,032	−3,365	−2,015	−2,015	−1,476	1,476	2,015	2,571	3,365	4,032	5,893
6	−5,208	−3,707	−3,143	−1,943	−1,943	−1,440	1,440	1,943	2,447	3,143	3,707	5,208
7	−4,785	−3,499	−2,998	−1,895	−1,895	−1,415	1,415	1,895	2,365	2,998	3,499	4,785
8	−4,501	−3,355	−2,896	−1,860	−1,860	−1,397	1,397	1,860	2,306	2,896	3,355	4,501
9	−4,297	−3,250	−2,821	−1,833	−1,833	−1,383	1,383	1,833	2,262	2,821	3,250	4,297
10	−4,144	−3,169	−2,764	−1,812	−1,812	−1,372	1,372	1,812	2,228	2,764	3,169	4,144
11	−4,025	−3,106	−2,718	−1,796	−1,796	−1,363	1,363	1,796	2,201	2,718	3,106	4,025
12	−3,930	−3,055	−2,681	−1,782	−1,782	−1,356	1,356	1,782	2,179	2,681	3,055	3,930
13	−3,852	−3,012	−2,650	−1,771	−1,771	−1,350	1,350	1,771	2,160	2,650	3,012	3,852
14	−3,787	−2,977	−2,624	−1,761	−1,761	−1,345	1,345	1,761	2,145	2,624	2,977	3,787
15	−3,733	−2,947	−2,602	−1,753	−1,753	−1,341	1,341	1,753	2,131	2,602	2,947	3,733
16	−3,686	−2,921	−2,583	−1,746	−1,746	−1,337	1,337	1,746	2,120	2,583	2,921	3,686
17	−3,646	−2,898	−2,567	−1,740	−1,740	−1,333	1,333	1,740	2,110	2,567	2,898	3,646
18	−3,610	−2,878	−2,552	−1,734	−1,734	−1,330	1,330	1,734	2,101	2,552	2,878	3,610
19	−3,579	−2,861	−2,539	−1,729	−1,729	−1,328	1,328	1,729	2,093	2,539	2,861	3,579
20	−3,552	−2,845	−2,528	−1,725	−1,725	−1,325	1,325	1,725	2,086	2,528	2,845	3,552
21	−3,505	−2,831	−2,518	−1,721	−1,721	−1,323	1,323	1,721	2,080	2,518	2,831	3,505
22	−3,505	−2,819	−2,508	−1,717	−1,717	−1,321	1,321	1,717	2,074	2,508	2,819	3,505
23	−3,485	−2,807	−2,500	−1,714	−1,714	−1,319	1,319	1,714	2,069	2,500	2,807	3,485
24	−3,467	−2,797	−2,192	−1,711	−1,711	−1,318	1,318	1,711	2,064	2,192	2,797	3,467
25	−3,450	−2,787	−2,485	−1,708	−1,708	−1,316	1,316	1,708	2,060	2,485	2,787	3,450
26	−3,435	−2,779	−2,479	−1,706	−1,706	−1,315	1,315	1,706	2,056	2,479	2,779	3,435
27	−3,421	−2,771	−2,473	−1,703	−1,703	−1,314	1,314	1,703	2,052	2,473	2,771	3,421
28	−3,408	−2,763	−2,467	−1,701	−1,701	−1,313	1,313	1,701	2,048	2,467	2,763	3,408
29	−3,396	−2,756	−2,462	−1,699	−1,699	−1,311	1,311	1,699	2,045	2,462	2,756	3,396
30	−3,385	−2,750	−2,457	−1,697	−1,697	−1,310	1,310	1,697	2,042	2,457	2,750	3,385
40	−3,307	−2,704	−2,423	−1,684	−1,684	−1,303	1,303	1,684	2,021	2,423	2,704	3,307
50	−3,261	−2,678	−2,403	−1,676	−1,676	−1,298	1,298	1,676	2,009	2,403	2,678	3,261
60	−3,232	−2,660	−2,390	−1,671	−1,671	−1,296	1,296	1,671	2,000	2,390	2,660	3,232
70	−3,211	−2,648	−2,381	−1,667	−1,667	−1,294	1,294	1,667	1,994	2,381	2,648	3,211
80	−3,195	−2,639	−2,374	−1,664	−1,664	−1,292	1,292	1,664	1,990	2,374	2,639	3,195
90	−3,183	−2,632	−2,369	−1,662	−1,662	−1,290	1,290	1,662	1,986	2,369	2,632	3,183
100	−3,174	−2,626	−2,365	−1,660	−1,660	−1,290	1,290	1,660	1,984	2,365	2,626	3,174
200	−3,131	−2,601	−2,345	−1,653	−1,653	−1,286	1,286	1,653	1,972	2,345	2,601	3,131
500	−3,092	−2,586	−2,334	−1,648	−1,648	−1,283	1,283	1,648	1,965	2,334	2,586	3,092
∞	−3,090	−2,576	−2,326	−1,645	−1,645	−1,282	1,282	1,645	1,960	2,326	2,576	3,090

Tabla F

Transformación *Z* de Fisher

Valores Z correspondientes a R_{XY} y ρ_{XY}

$R_{XY} \mid \rho_{XY}$	Z	$R_{XY} \mid \rho_{XY}$	Z	$R_{XY} \mid \rho_{XY}$	Z	$R_{XY} \mid \rho_{XY}$	Z
0,00	0,0000	0,25	0,2554	0,50	0,5493	0,75	0,9730
0,01	0,0100	0,26	0,2661	0,51	0,5627	0,76	0,9962
0,02	0,0200	0,27	0,2769	0,52	0,5763	0,77	1,0203
0,03	0,0300	0,28	0,2877	0,53	0,5901	0,78	1,0454
0,04	0,0400	0,29	0,2986	0,54	0,6042	0,79	1,0714
0,05	0,0500	0,30	0,3095	0,55	0,6184	0,80	1,0986
0,06	0,0601	0,31	0,3205	0,56	0,6328	0,81	1,1270
0,07	0,0701	0,32	0,3316	0,57	0,6475	0,82	1,1568
0,08	0,0802	0,33	0,3428	0,58	0,6625	0,83	1,1881
0,09	0,0902	0,34	0,3541	0,59	0,6777	0,84	1,2212
0,10	0,1003	0,35	0,3654	0,60	0,6931	0,85	1,2562
0,11	0,1104	0,36	0,3769	0,61	0,7089	0,86	1,2933
0,12	0,1206	0,37	0,3884	0,62	0,7250	0,87	1,3331
0,13	0,1307	0,38	0,4001	0,63	0,7414	0,88	1,3758
0,14	0,1409	0,39	0,4118	0,64	0,7582	0,89	1,4219
0,15	0,1511	0,40	0,4236	0,65	0,7753	0,90	1,4722
0,16	0,1614	0,41	0,4356	0,66	0,7928	0,91	1,5275
0,17	0,1717	0,42	0,4477	0,67	0,8107	0,92	1,5890
0,18	0,1820	0,43	0,4599	0,68	0,8291	0,93	1,6584
0,19	0,1923	0,44	0,4722	0,69	0,8480	0,94	1,7380
0,20	0,2027	0,45	0,4847	0,70	0,8673	0,95	1,8318
0,21	0,2132	0,46	0,4973	0,71	0,8872	0,96	1,9459
0,22	0,2237	0,47	0,5101	0,72	0,9076	0,97	2,0923
0,23	0,2342	0,48	0,5230	0,73	0,9287	0,98	2,2976
0,24	0,2448	0,49	0,5361	0,74	0,9505	0,99	2,6467

Glosario de símbolos

A_{IQ}	amplitud intercuartil.
A_T	amplitud total.
$B(n, \pi_1)$	distribución teórica binomial, con parámetros n y π_1.
C	coeficiente de contingencia.
$C_{\text{máx}}$	valor máximo del coeficiente de contingencia.
C_k	centiles.
$C_{N,n}$	combinaciones sin repetición de N elementos tomados de n en n.
CV_{media}	coeficiente de variación centrado en la media.
CV_{mediana}	coeficiente de variación centrado en la mediana.
D_k	deciles.
E	espacio muestral.
$E_{\text{máx}}$	error máximo en los intervalos de confianza.
$E(Y)$	valor esperado de la variable Y.
$f(Y)$	función de probabilidad (o de densidad) de la variable Y.
$F(Y)$	función de probabilidad (o de densidad) acumulada de la variable Y.
gl	grados de libertad.
g_1	índice de asimetría.
g_2	índice de curtosis.
H_0	hipótesis nula en los contrastes de hipótesis.
H_1	hipótesis alternativa en los contrastes de hipótesis.
i	i-ésimo valor de una variable. En variables categóricas: $i = 1, 2, ..., I$. En variables cuantitativas: $i = 1, 2, ..., n$.
j	j-ésimo valor de una variable categórica: $j = 1, 2, ..., J$. También, j-ésimo grupo.
IC_θ	intervalo de confianza para el parámetro θ.
IVC	índice de variación cualitativa.
L_i	límite inferior de un intervalo de confianza.
L_s	límite superior de un intervalo de confianza.
m_i	frecuencias teóricas o esperadas en una tabla de contingencias unidimensional.
m_{ij}	frecuencias teóricas o esperadas en una tabla de contingencias bidimensional.
$Mdn_{\text{|desv|}}$	mediana de las desviaciones.
Mdn_Y	mediana de la variable Y.
$M(n, \pi_i)$	distribución teórica multinomial, con parámetros n y π_i.
M_r	momento de orden r.
n	número de casos en la muestra.

N	número de casos en la población.
$N(a, b)$	distribución teórica normal, con parámetros a y b.
na_i	frecuencia absoluta acumulada.
n_i	frecuencias absolutas. Variable multinomial.
n_j	número de casos en el grupo j (j = 1, 2, ..., J).
n_{ij}	frecuencias conjuntas en una tabla de contingencias.
n_{i+}	frecuencias marginales de las filas en una tabla de contingencias.
n_{+j}	frecuencias marginales de las columnas en una tabla de contingencias.
n_1	número de “unos” (éxitos). Variable binomial.
p	nivel crítico.
P_1	proporción de “unos” (éxitos). Variable binomial.
Pa_i	frecuencia relativa acumulada.
P_i	frecuencia relativa. Variable multinomial.
P_k	percentiles.
P_n	permutaciones sin repetición.
$P(S)$	probabilidad de un suceso.
$P(S_1 \cup S_2)$	probabilidad de la unión de dos sucesos.
$P(S_1 \cap S_2)$	probabilidad de la intersección de dos sucesos.
$P(S_1 \mid S_2)$	probabilidad condicional.
$\bar{Q}$	trimedia.
Q_k	cuartiles.
R_i	residuos en una tabla de contingencias unidimensional.
R_{ij}	residuos en una tabla de contingencias bidimensional.
R_{XY}	coeficiente de correlación de Pearson entre las variables X e Y.
S_{g_1}	error típico del índice de asimetría.
S_{g_2}	error típico del índice de curtosis.
S_Y	desviación típica de la variable Y (siempre la insesgada).
S_Y^2	varianza de la variable Y (siempre la insesgada).
S_{XY}	covarianza entre las variables X e Y.
t_{gl}	distribución teórica t de Student con gl grados de libertad.
T	variable distribuida según el modelo teórico de probabilidad t de Student.
V	varianza de una variable.
$V_{\text{cramèr}}$	coeficiente V de Cramèr.
$V_{N,n}$	variaciones sin repetición de N elementos tomados de n en n.
X^2	estadístico de Pearson distribuido según el modelo de probabilidad ji-cuadrado.
y	puntuaciones diferenciales de la variable Y.
y_i	puntuaciones diferenciales de la variable Y.
Y	puntuaciones directas de la variable Y.
Y_i	puntuaciones directas de la variable Y.
$\bar{Y}$	media de la variable Y.
$\bar{Y}_{\text{arm}}$	media armónica de la variable Y.

$\overline{Y}_{|desv|}$ media de las desviaciones.
$\overline{Y}_{geom}$ media geométrica de la variable *Y*.
$\overline{Y}_j$ media de la variable *Y* en el *j*-ésimo grupo.
$\overline{Y}_W$ media *winsorizada* de la variable *Y*.
$\overline{Y}_{k\,\%}$ media recortada o truncada de la variable *Y*.
Z puntuaciones típicas.
Z_i puntuaciones típicas.
Z_{R_i} residuos tipificados.
$Z_{R_{XY}}$ transformación *Z* de Fisher para el coeficiente de correlación de Pearson.
$\%a_i$ frecuencia porcentual acumulada.
$\%_i$ frecuencia porcentual.

Letras griegas utilizadas

α nivel de significación o riesgo en los contrastes de hipótesis y en los intervalos de confianza.
θ forma genérica de identificar un parámetro.
$\hat{\theta}$ forma genérica de identificar un estadístico utilizado como estimador.
μ_Y valor esperado (media poblacional o teórica) de la variable *Y*.
ν notación genérica para los grados de libertad.
π_1 proporción teórica en una variable dicotómica.
π_i proporción teórica en una variable categórica.
π_{ij} proporción teórica en una tabla de contingencias bidimensional.
π_{i+} proporción teórica marginal en una tabla de contingencias bidimensional.
π_{+j} proporción teórica marginal en una tabla de contingencias bidimensional.
Π símbolo del producto.
ρ_{XY} coeficiente de correlación de Pearson en la población.
σ_Y^2 varianza teórica o poblacional de la variable *Y*.
σ_Y desviación típica teórica o poblacional de la variable *Y*.
Σ símbolo del sumatorio.
φ coeficiente de correlación "phi".
χ^2 distribución teórica de probabilidad *ji*-cuadrado.
χ_{gl}^2 distribución teórica de probabilidad *ji*-cuadrado con *gl* grados de libertad.

Referencias

Agresti A (2002). *Categorical data analysis* (2ª ed). New York: Wiley.

Agresti A (2007). *An introduction to categorical data analysis* (2ª ed). New York: Wiley.

Agresti A y Finlay B (1986). *Statistical methods for the social sciences* (2ª ed). San Francisco, CA: Dellen.

Amón J (1979). *Estadística para psicólogos*. Vol 1: *Estadística descriptiva* (2ª ed). Madrid: Pirámide.

Amón J (1984). *Estadística para psicólogos*. Vol 2: *Probabilidad y estadística inferencial* (3ª ed). Madrid: Pirámide.

Anderson NH (1961). Scales and statistics: Parametric and nonparametric. *Psychological Bulletin*, *58*, 305-316.

Andrews DF, Bickel PJ, Hampel FR, Huber PJ, Rogers WH y Tukey JW (1972). *Robust estimates of location. Survey and advances*. Princeton, NJ: Princeton University Press.

Argyrous G (2005). *Statistics for research, with a guide for SPSS* (2ª ed). London: Sage.

Beherens JT (1997). Principles and procedures of exploratory data analysis. *Psychological Methods*, *2*, 131-160.

Binder A (1984). Restrictions on statistics imposed by method of measurement: Some reality, much mythology. *Journal of Criminal Justice*, *12*, 467-481.

Bliss CI (1967). *Statistics in biology* (Vol. 1). Nueva York: McGraw-Hill.

Boneau CA (1960). Test of violations of assumptions underlying the t test. *Psychological Bulletin*, *57*, 49-64.

Boss DD y Hughes-Oliver JM (2000). How large does n have to be for the Z and t intervals. *The American Statistician*, *54*, 121-128.

Botella J, León OG, San Martín R y Barriopedro MI (2001). *Análisis de datos en psicología I. Teoría y ejercicios*. Madrid: Pirámide.

Bradley DR, Bradley TD, McGrath SG y Cutcomb SD (1979). Type I error rate of the chi-square test of independence in $R \times C$ tables that have small expected frequencies. *Psychological Bulletin*, *86*, 1290-1297.

Bryk AS y Raudenbush SW (1988). Heterogeneity of variance in experimental studies: A challenge to conventional interpretations. *Psychological Bulletin*, *104*, 396-404.

Camili G y Hopkins KD (1978). Applicability of chi-square to 2×2 contingency tables with small expected cell frequencies. *Psychological Bulletin*, *85*, 163-167.

Camili G y Hopkins KD (1979). Testing for association in 2×2 contingency tables with very small sample sizes. *Psychological Bulletin*, *86*, 1011-1014.

Chow SL (1996). *Statistical significance: Rationale, validity and utility*. Thousand Oaks, CA: Sage.

Cochran WG (1952). The χ^2 test of goodness of fit. *Annals of Mathematical Statistics*, *23*, 315-345.

Cohen, J (1990). Things I have learned (so far). *American Psychologist*, *45*, 1304-1312.

Cohen, J (1994). The earth is round ($p < .05$). *American Psychologist*, *49*, 997-1003.

Conover WJ, Johnson ME y Johnson MM (1981). A comparative study of test for homogeneity of variances with applications to the outer continental self bidding data. *Technometrics, 23*, 351-361.

Cortina JM y Dunlap WP (1997). On the logic and purpose of significance testing. *Psychological Methods*, *2*, 171-172.

Cramèr H (1946). *Mathematical methods of statistics*. Princeton, NJ: Princeton University Press.

Davis JA (1985). *The logic of causal order*. Beverly Hills, CA: Sage.

Davison ML y Sharma AR (1988). Parametric statistics and levels of measurement. *Psychological Bulletin*, 1988, *104*, 137-144.

Dunbar G (2005). *Evaluating research methods in psychology. A case study approach*. Malden, MA: BPS Blackwell.

Field A (2009). *Discovering statistics using SPSS* (3ª ed). London: Sage.

Fisher RA (1918). The correlation between relations on the supposition of mendelian inheritance. *Transactions of the Royal Society of Edinburgh*, LII.

Fisher RA (1921). On the probable error of a coeficient correlation deduced from a small sample. *Metron, 1,* 3,32.

Fisher RA (1925). *Statistical methods for research workers*. Edinburg: Oliver and Boyd.

Fisher RA (1955). Statistical methods and scientific induction. *Journal of the Royal Statistical Society, 17,* 69-78.

Foster JJ (2001). *Data analysis using SPSS for Windows. A beginner's guide* (2ª ed). London: Sage.

Gaito J (1960). Scale classifications and statistics. *Psychological Review*, *67*, 277-278.

Gaito J (1980). Measurement scales and statistics: Resurgence of an old misconception. *Psychological Bulletin*, *87*, 564-567.

Gaito J (1986). Some issues in the measurement-statistics controversy. *Canadian Psychology*, *27*, 63- 68.

Gardner PL (1975). Scales and statistics. *Review of Educational Research*, *45*, 43-57.

Gigerenzer G (1993). The superego, the ego, and the id in statistical reasoning. En G Keren y C Lewis (Eds), *A handbook for data analysis in the behavioral sciences. Methodological issues* (págs. 311-339). Hillsdale, NJ: Lawrence Erlbaum Associates.

Gosset WS (Student) (1908). The probable error of a mean. *Biometrika*, *6*, 1-25.

Haberman SJ (1973). The analysis of residuals in cross-classification tables. *Biometrics*, *29*, 205-220.

Hagen RL (1997). In praise of the hypothesis statistical test. *American Psychologist*, *52*, 15-24.

Hardy M y Briman A (Eds) (2004). *Handbook of data analysis*. London: Sage.

Harlow LL, Mulaik SA y Steiger JH (1997). *What if there were no significance test*. Mahwah, NJ: Lawrence Erlbaum Associates.

Hartley HO (1940). Testing the homogeneity of a set of variances. *Biometrika*, *31*, 249-255.

Hartley HO (1950). The maximum F-ratio as a short-cut test for heterogeneity of variance. *Biometrika*, *37*, 308-312.

Hays WL (1994). *Statistics* (5ª ed). New York: Holt, Rinehart and Winston.

Hill AB (1965). The environment and disease: Association o causation? *Proceedings of the Royal Society of Medicine*, *58*, 295-300.

Hoaglin DC, Mosteller F y Tukey JW (Eds) (1983). *Understanding robust and exploratory data analysis*. New York: Wiley.

Höfer Th, Przyrembel H y Verleger S (2004). New evidence for the theory of the stork. *Paediatric and Perinatal Epidemiology*, *18*, 88-92.

Hogg RV y Craig AT (1970). *Introduction to mathematical statistics* (3ª ed). New York: McMillan.

Hotelling H (1931). The generalization of Student's ratio. *Annals of Mathematical Statistics*, *2*, 360-378.

Howell DC (2002). *Statistical methods for psychology* (5ª ed). Belmont, CA: Thomson Wadsworth.

James GS (1951). The comparison of several groups of observations when the ratios of the population variances are unknown. *Biometrika, 38*, 324–329.

Kaplan D (Ed) (2004). *The Sage handbook of quantitative methodology for the social sciences*. Thousand Oaks, CA: Sage.

Kenny DA y Judd ChM (1986). Consequences of violating the independence assumption in analysis of variance. *Psychological Bulletin, 99,* 422-431.

Keppel G y Wickens ThD (2004). *Design and analysis. A researcher's handbook* (4ª ed). Englewood Cliffs, NJ: Prentice-Hall.

Keren G y Lewis C (Eds) (1993). *A handbook for data analysis in the behavioral sciences. Methodological issues*. Hillsdale, NJ: Lawrence Erlbaum Associates.
Keren G y Lewis C (Eds) (1993). *A handbook for data analysis in the behavioral sciences. Statistical issues*. Hillsdale, NJ: Lawrence Erlbaum Associates.
Kinnear PR y Gray CD (2000). *SPSS for Windows made simple*. Hove: Psychology Press.
Kirk RE (1978). *Introductory statistics*. Monterey, CA: Wadsworth.
Kirk RE (1996). Practical significance. A concept whose time has come. *Educational and Psychological measurement*, *56*, 746-759.
Kirk RE (1999). *Statistics. An introduction*. London: Harcourt Brace.
Kornbrot DE (1990). The rank difference test: A new and meaningful alternative to the Wilcoxon signed ranks test for ordinal data. *British Journal of Mathematical and Statistical Psychology*, *43*, 241-261.
Landau S y Everitt B (2004). *A handbook of statistical analisys using SPSS*. Boca Raton, FL: Chapman and Hall.
Larntz K (1978). Small-sample comparison of exact levels for chi-square goodness-of-fit statistics. *Journal of the American Statistical Association*, *73*, 25-263.
León OG (1984). El uso del término "significativo" en los informes experimentales. *Revista de Psicología General y Aplicada*, *39*, 455-469.
León OG y Montero I (2003). *Métodos de investigación en psicología y educación* (3ª ed). Madrid: McGraw-Hill.
Levene H (1960). Robust tests for the equality of variances. En J. Olkin (Ed), *Contributions to probability and statistics*. Palo Alto, CA: Stanford University Press.
Lewis PAW y Orav EJ (1989). *Simulation methodology for statisticians, operation analysts and engineers* . Belmont, CA: Wadsworth.
Linoff GS (2008). *Data Analysis using SQL and Excel*. Indianapolis, IN: Wiley.
Maxwell SE y Delaney HD (1985). Measurement and statistics: An examination of construct validity. *Psychological Bulletin*, *97*, 85-93.
Maxwell SE y Delaney HD (2004). *Designing experiments and analyzing data* (2ª ed). Mahwah, NJ: Lawrence Erlbaum Associates.
Miller RG (1986). *Beyond ANOVA. Basics of applied statistics*. New York: Wiley.
Moore DS (2007). *The basic practice of statistics* (4ª ed). New York: WH Freeman and Company.
Moore DS y Notz WI (2006). *Statistics. Concepts and controversies* (6ª ed). New York: WH Freeman and Company.
Morrison DE y Henkel RE (Eds) (1970). *The significant test controversy: A reader*. Chicago, IL: Aldine.
Mueller JH y Schuessler KF (1961). *Statistical reasoning in sociology*. Boston: Houghton Mifflin.
Neyman J y Pearson ES (1928). On the use and interpretation of certain test criteria for purposes of statistical inference (2ª parte). *Biometrika*, *20*, 263-294.
Neyman J y Pearson ES (1932). The testing of statistical hypotheses in relation to probabilities a priori. *Proceedings of the Cambridge Philosophical Society, 29,* 492-516.
Neyman J y Pearson ES (1933). On the problem of the most efficient test of statistical hipotheses. *Philosophical Transactions of the Royal Society, 231,* 284-337.
Norušis MJ y SPSS Inc (1993). *SPSS for Windows. Base system user's guide release 6.0*. Chicago, IL: SPSS Inc.
Norušis MJ (2003). *SPSS 12.0. Statistical procedures companion*. Upper Saddle River, NJ: Prentice Hall.
Oakes M (1986). *Statistical inference. A commentary for the social and behavioral sciences*. New York: Wiley.
O'Brien RG (1981). A simple test for variance effects in experimental designs. *Psychological Bulletin*, *89*, 570-574.
Overall JE (1980). Power of chi-square test for 2×2 contingency tables with small expected frequencies. *Psychological Bulletin*, *87*, 132-135.

Palmer AL (1999). *Análisis de datos. Etapa exploratoria*. Madrid: Pirámide.

Pardo A (2002). *Análisis de datos categóricos*. Madrid: UNED.

Pardo A y Alonso J (1990). *Motivar en el aula*. Madrid: Ediciones de la Universidad Autónoma.

Pardo A y Ruiz MA R (2009). *Gestión de datos con SPSS Statistics*. Madrid: Síntesis.

Pardo A y San Martín R (1994). *Análisis de datos en psicología II*. Madrid: Pirámide.

Pardo A y San Martín R (1998). *Análisis de datos en psicología II* (2ª ed). Madrid: Pirámide.

Pearson ES y Please NW (1975). Relation between the shape of population distribution and the robustness of four simple test statistics. *Biometrika*, *62*, 223-241.

Pearson K (1896). Mathematical contributions to the theory of evolution. III. Regression, heredity, and panmixia. *Philosophical Transactions of the Royal Society*, *A*, *757*, 253-318.

Pearson K (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can reasonably be supposed to have arisen from random sampling. *Philosophical Magazine*, *50*, 157-175.

Pearson K (1911). On the probability that two independent distributions of frequency are really samples from the same population. *Biometrika*, *8*, 250-254.

Pearson K (1913). On the probable error of a correlation coefficient as found from a fourfold table. *Biometrika*, *9*, 22-27.

Pereda S (1987). *Psicología experimental. 1: Metodología*. Madrid: Pirámide.

Posten HO (1978). The robustness of the two-sample *t*-test over the Pearson system. *Journal of Statistical Computation and Simulation*, *6*, 295-311.

Posten HO (1979). The robustness of the one-sample *t*-test over the Pearson system. *Journal of Statistical Computation and Simulation*, *9*, 133-149.

Ramsay JO y Silverman BW (2002). *Applied functional data analysis. Methods and case studies*. New York: Springer-Verlag.

Ramsey P (1980). Exact type I error rates for robustness of Student's *t* test with unequal variances. *Journal of Educational Statistics*, *5*, 337-349.

Ríos S (1985). *Métodos estadísticos* (2ª ed, 3ª reimp) Madrid: Ediciones del Castillo.

Rosenthal R y Rosnow RL (1991). *Essentials of behavioral research. Methods and data analysis* (2ª ed). New York: McGraw Hill.

Rosnow RL y Rosenthal R (2005). *Beginning behavioral research. A conceptual primer* (5ª ed). Englewood Cliffs, NJ: Prentice Hall.

Ruscio J (2006). *Critical thinking in psychology: Separating sense of nonsense* (2ª ed). Belmont, CA: Thomson Wadsworth.

San Martín R y Pardo A (1989). *Psicoestadística. Contrastes paramétricos y no paramétricos*. Madrid: Pirámide.

Satterthwaite FE (1946). An approximate distribution of estimates of variance components. *Biometrics*, *2*, 110-114.

Saville BK (2008). *Research methods in psychology. A guide to teaching*. Malden, MA: Blackwell Publishing.

Sawilowsky S y Blair RC (1992). A more realistic look at the robustness and type II error properties of the *t* test to departures from population normality. *Psychological Bulletin*, *111*, 352-360.

Schay G (2007). Introduction to probability with statistical applications. Berlín: Birkhäuser.

Schmidt, FL (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training research. *Psychological Methods*, *1*, 115-129.

Siegel S y Castellan NJ (1988). *Nonparametric statistics for the behavioral sciences* (2ª ed). New York: McGraw-Hill.

Sies H (1988). A new parameter for sex education. *Nature*, *332*, 495.

Solanas A, Salafranca L, Fauquet J y Núñez MI (2005). *Estadística descriptiva en ciencias del comportamiento*. Madrid: Thomson.

Steiger JH (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, *87*, 245-251.

Stevens SS (1946). On the theory of scales of measurement. *Science*, *103*, 677-680.

Stevens SS (1951). Mathematics, measurement and psychophysics. En SS Stevens (Ed), *Handbook of experimental psychology* (págs. 1-49). New York: Wiley.

Tamhane AC y Dunlop D (2000). *Statistics and data analysis. From elementary to intermediate*.Upper Saddle River, NJ: Prentice Hall.

Townsend JT y Ashby FG (1984). Measurement scales and statistics: The misconception misconceived. *Psychological Bulletin*, *96*, 394-401.

Tukey JW (1962). The future of data analysis. *Annals of Mathematical Statistics*, *33*, 1-67.

Tukey JW (1977). *Exploratory data analysis*. Reading, MA: Addison Wesley.

Velleman PF y Hoaglin DC (2004). *Applications, basics and computing of exploratory data analysis*. Ithaca, NY: The Internet-First University Press (Universidad de Cornell).

Wassertheil-Smoller S (2004). *Biostatistics and epidemiology. A primer for health and biomedical professionals*. New York: Springer-Verlag.

Welch BL (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika, 29,* 350-362.

Welch BL (1947). The generalization of Student's problem when several difference population variances are involved. *Biometrika, 34,* 29-35.

Westerman R (1983). Interval-scale measurement of attitudes: Some theoretical and empirical testing methods. *British Journal of Mathematical and Statistical Psychology*, *36*, 228- 239.

Wickens ThD (1989). *Multiway contingency tables analysis for the social sciences*. Hillsdale, NJ: LEA.

Wilcox RR y Keselman HJ (2003). Modern robust data analysis methods: Measures of central tendency. *Psychological Methods*, *8*, 254–274.

Williams EJ (1959). The comparison of regression variables. *Journal of the Royal Statistical Society*, *Series B*, *21*, 396-399.

Winer BJ, Brown DR y Michels KM (1991). *Statistical principles in experimental design* (3ª ed). New York: McGraw-Hill.

Wonnacott TH y Wonnacott RJ (1990). *Introductory statistics* (5ª ed). New York: Wiley.

Wright DB (1997a). Football standings and measurement levels. *Journal of The Royal Statistical Society*, *Series D* (*The Statistician*), *46*, 105-110.

Wright DB (1997b). *Understanding statistics. An introduction for the social sciences*. London: Sage.

Wright DB (2002). *First steps in statistics*. London: Sage.

Wright DB (2002). Making friends with your data: Improving how statistics are conducted and reported. *British Journal of Educational Psychology*, *73*, 123-136.

Zou GY (2007). Toward using confidence intervals to compare correlations. *Psychological Methods*, *12*, 399-413

Zumbo BD y Zimmerman DW (2000). Scales of measurement and the relation between parametric and nonparametric statistical tests. En B Thomson (Ed), *Advances in social science methodology* (Vol 6). Greenwich, CT: JAI Press.

Zuur AK, Ieno EN y Smith GM (2007). *Analyzing Ecological Data*. New York: Springer-Verlag.

Índice de materias

D

E

L

M

N

O

P